UNIVERSITY OF CALIFORNIA SAN DIEGO

Statistical and Computational Methods for Analyzing Accelerometer Data

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics

by

Yue Xu

Committee in charge:

Professor Ian Abramson, Co-Chair
Professor Loki Natarajan, Co-Chair
Professor Ery Arias-Castro
Professor Michael Donohue
Professor Jacqueline Kerr
Professor Dimitris N. Politis

2018

The dissertation of Yue Xu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California San Diego

2018

DEDICATION

To Google, without which this paper would be completed two years later.

EPIGRAPH

All models are wrong

but some are useful.

—George Box

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my two amazing advisors Professor Ian Abramson and Professor Loki Natarajan. Ian put this super team together three years ago and has since offered me useful advice on theory development and refinement for my PhD study and related research. I'm thankful for his patience, motivation, and immense knowledge. Loki has given me invaluable guidance that helped me through every step of my research study from conceptualizing the problem to writing of this thesis. I am indebted for her constant assistance, encouragement, and the tremendous support and opportunities she provided throughout my doctoral studies at UC San Diego. I could not have imagined having a better advisor and mentor for my PhD study.

Besides my advisors, I would like to thank the rest of my thesis committee: Professor Ery Arias-Castro, Professor Dimitris N. Politis, Professor Jacqueline Kerr, and Professor Michael Donohue, not only for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

Chapter 1, in full, is a reprint of the material as it appears in "Statistical approaches to account for missing values in accelerometer data: Applications to modeling physical activity", *Statistical Methods in Medical Research*, Vol 27, Issue 4, pp.1168-1186. Selene Yue Xu, Sandahl Nelson, Jacqueline Kerr, Suneeta Godbole, Ruth Patterson, Gina Merchant, Ian Abramson, John Staudenmayer, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in "Cognition, qualityoflife, and symptom clusters in breast cancer: Using Bayesian networks to elucidate complex relationships", *PsychoOncology*, 2017. Selene Yue Xu, Wesley Thompson, Sonia Ancoli-Israel, Lianqi Liu, Barton Palmer, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it is under review in "Modeling interre-

lationships between health behaviors: applying Bayesian networks". Selene Yue Xu, Wesley Thompson, Jacqueline Kerr, Suneeta Godbole, Dorothy Sears, Ruth Patterson, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it is under review in "Using quasi-Poisson processes to model accelerometer data". Selene Yue Xu, Jacqueline Kerr, Suneeta Godbole, Ruth E Patterson, Cheryl L Rock, Ian Abramson, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout my academic pursuit.

VITA

2009-2012	B. A. in Statistics and B. A. in Economics, University of California Berkeley
2013-2018	Ph. D. in Mathematics with Specialization in Statistics, University of California San Diego

PUBLICATIONS

Selene Yue Xu, Sandahl Nelson, Jacqueline Kerr, Suneeta Godbole, Ruth Patterson, Gina Merchant, Ian Abramson, John Staudenmayer, Loki Natarajan, "Statistical approaches to account for missing values in accelerometer data: Applications to modeling physical activity", *Statistical Methods in Medical Research*, Vol 27, Issue 4, pp.1168-1186.

Selene Yue Xu, Wesley Thompson, Sonia Ancoli-Israel, Lianqi Liu, Barton Palmer, and Loki Natarajan, "Cognition, qualityoflife, and symptom clusters in breast cancer: Using Bayesian networks to elucidate complex relationships", *PsychoOncology*, 2017.

Selene Yue Xu, Sandahl Nelson, Jacqueline Kerr, Suneeta Godbole, Eileen Johnson, Ruth E Patterson, Cheryl L Rock, Dorothy D Sears, Ian Abramson, Loki Natarajan, "Modeling temporal variation in physical activity using functional principal components analysis", *under review*

Selene Yue Xu, Wesley Thompson, Jacqueline Kerr, Suneeta Godbole, Dorothy Sears, Ruth Patterson, Loki Natarajan, "Modeling interrelationships between health behaviors: applying Bayesian networks", *under review*

Selene Yue Xu, Jacqueline Kerr, Suneeta Godbole, Ruth E Patterson, Cheryl L Rock, Ian Abramson, Loki Natarajan, "Using quasi-Poisson processes to model accelerometer data", *under review*

Selene Yue Xu, Sandahl Nelson, Jacqueline Kerr, Suneeta Godbole, Ruth Patterson, Ian Abramson, Loki Natarajan, (2015) Objective assessment of physical activity: statistical approaches to account for missing values in accelerometry data, Poster presentation, UC San Diego Public Health Research Day, San Diego, CA. Won third place in Poster competition (graduate category)

Selene Yue Xu, Wesley Thompson, Jacqueline Kerr, Suneeta Godbole, Ruth Patterson, Loki Natarajan, (2016) Modeling interrelationships between lifestyle behaviors using Bayesian networks, Poster presentation, UC San Diego Public Health Research Day, San Diego, CA. Won first place in Poster competition (graduate category)

Selene Yue Xu, Wesley Thompson, Jacqueline Kerr, Suneeta Godbole, Ruth Patterson, Loki Natarajan, (2017) Modeling interrelationships between lifestyle behaviors using Bayesian networks, Poster presentation, One-day ASA conference, San Diego, Won Best Poster award

ABSTRACT OF THE DISSERTATION

Statistical and Computational Methods for Analyzing Accelerometer Data

by

Yue Xu

Doctor of Philosophy in Mathematics

University of California San Diego, 2018

Professor Ian Abramson, Co-Chair
Professor Loki Natarajan, Co-Chair

Advances in technology have resulted in the use of sensors in a great variety of applications ranging from weather forecasting, GPS tracking to physical activity measurement. Novel analytic techniques are being developed to study these densely sampled data. My research projects focus on approaches to analyze and model accelerometry data. Accelerometers measure minute-level human movement, and hence provide a rich framework for assessing physical activity patterns of an individual. Using accelerometer data collected in research studies in the School of Medicine at the University of California San Diego, our objectives are (a) to ascertain activity patterns incorporating temporal and subject-to-subject variation (b) to test if these patterns are associated

with health outcomes such as obesity, cancer status, biomarkers and quality of life. We apply modern machine learning techniques and develop novel mathematical frameworks to analyze these big data. We anticipate that this work will provide statistical and computational tools to study accelerometry and inform societal guidelines on leading a healthy lifestyle.

Chapter 1

Missing Data Project

1.1 Introduction

Physical inactivity and sedentary behavior are recognized risk factors for many chronic diseases [BCM94] [SMG+05] [Bla09] [WEA+12], driving research on levels of physical activity needed to maintain a healthy lifestyle and prevent disease. Accelerometers are objective means to measure duration and intensity of daily physical activity, and may be less prone to the biases associated with self-report [SS00] [PAH+08]. Traditional protocols for hip-worn accelerometers instruct participants to wear the device for at least 5 days during waking hours [WEV+05]. Unfortunately missing data, due to participants removing their accelerometer for varying and undocumented reasons, leads to non-random bias, which in turn often results in inaccurate assessments of physical activity. These missing data and attending biases present a major obstacle to the interpretation of accelerometer-based research.

Previous studies have highlighted the error caused by inconsistencies in the number of wear days across participants [KBB+09]. In addition, errors due to variations in the amount of wear time each day have been outlined, with substantial bias noted when daily wear time was less than 12 hours [HBK+13] [HBK+14]. However, including only days with $\geq$ 12 hours of device wear

would result in researchers being forced to discard a large amount of otherwise usable participant data. To combat these potential biases and yet make optimal use of available information, the majority of accelerometer studies include only data that meets a minimum required number of days ($\geq$ 3-5 days [TMP05] [M$^+$15]) and time per day (ranging 6-10 hours/day [M$^+$15] [CCT10] [YJC$^+$09]), and then account for wear time variation by either (a) using imputation methods [SZC12] [CHM$^+$05] [WVHS$^+$15] (b) normalizing activity measures by wear time [YJC$^+$09] [KRB$^+$09] [KM14], and/or (c) adjusting for wear time in regression models [SFT$^+$13] [vLWV$^+$12]. In addition, a few use Bayesian techniques to incorporate individuals with as little as 1 day of valid wear time [CGJ$^+$11] [TBD$^+$08]. Thus, there is as yet no consensus regarding the optimal analytic method for accounting for non-wear time, with a variety of methods in use, making it difficult to compare results across studies.

In this article, we focused on a regression modeling framework, and aimed to develop and evaluate statistical methods to standardize analysis and accurately estimate regression parameters of interest despite the presence of missing data due to non-wear. A primary objective was to develop methods that would be easy to implement using standard software and thus accessible to the physical activity research community. We implemented a pseudosimulation approach, whereby realistic missing data patterns were simulated. We used baseline data from a cohort of postmenopausal breast cancer survivors to create the pseudo-simulated data sets. Next, a variety of statistical methods to account for variability in device wear time were applied to these simulated data; bias and precision of these methods were compared. The outline of this article is as follows: in Section 1.2, we provide details on demographics and accelerometer measures available for our study sample. In Sections 1.3 and 1.4, we specify the objectives of our analysis, what we consider "complete" profiles for the purposes of the simulations, and the regression model and parameters of interest. Section 1.5 covers our algorithm for simulating missing data patterns from complete profiles. In Sections 1.6 - 1.9 we propose three analytic methods to account for missing values in accelerometry data, evaluate the performances of these methods,

and discuss a Poisson framework to justify the relative success of one of our proposed methods. We conclude (Section 1.10) with a discussion of strengths and weaknesses of our approach and some recommendations for use of the methods.

1.2 Study Sample

Our study cohort comprised of 333 overweight postmenopausal breast cancer survivors participating in a weight-loss intervention trial [PMN$^+$15]. Participants were on average 63 (SD=6.9) years with mean BMI 31.1 (SD=4.9) at study entry; 51% had college degree or higher, 48% had Stage I, 35% Stage II, and 17% Stage III breast cancer.

Objective physical activity in our study was assessed via the GT3X Actigraph (ActiGraph, LLC; Pensacola, FL), which is a triaxial lightweight accelerometer approximately 2 X 2 X 1 in size. The Actigraph GT3X+ monitor was set to collect acceleration data at 30 Hz. The ActiLife program applied a band-pass filter to remove non-human acceleration signal from the data and then summarized the signal to counts per minute using a proprietary algorithm [Bas12b]. The magnitude of the count is related to intensity of the activity [Bas12b]. The device has been validated and calibrated for use in both controlled and field conditions [Bas12b]. Participants were provided with written protocols for best positioning of the device and instructed to wear it on the hip for 7 days during all waking hours, except for when in contact with water. Non-wear time was identified via pre-defined algorithms of consecutive zero counts using standard protocols [CLMB11] and labeled as missing data.

Our sample had a total of 2814 days of accelerometer data, with the number of days per participant ranging from 2 to 21. Median wear time per day was 816 minutes (25%-ile = 725 minutes, 75%-ile = 895 minutes).

1.3 Study Objectives and General Strategy

A major objective in behavioral studies is identifying demographic and other factors associated with a behavior of interest (e.g., dietary intake or physical activity level). Hence our primary objective in the current study was to fit regression models to identify factors associated with physical activity level.

For the first set of analyses, we used physical activity volume, i.e. total accelerometer counts per minute, as our measure of physical activity. In later sections (Section 1.8), we consider other measures (e.g., moderate-to-vigorous activity [MVPA], and sedentary time). Our data comprised accelerometer counts measured on multiple days for each participant. To account for these hierarchical data, we used linear mixed effects models [DHLZ13] with accelerometer count per minute as the nested dependent variable, and age, BMI, depression, and weekend status as independent predictors. Mathematically, our model was specified as:

$$Y_{ij} = \alpha_i + A_i\beta_1 + B_i\beta_2 + D_i\beta_3 + W_{ij}\beta_4 + \varepsilon_{ij} \tag{1.1}$$

Here Y_{ij} is the average activity count for the jth day of the ith subject. α_i is the random intercept for subject i. A_i is the age for the ith subject. B_i is the BMI for the ith subject. D_i is the depression indicator (1 if depressed, 0 if otherwise) for the ith subject. W_{ij} is the weekend indicator (1 if it is a weekend day, 0 if otherwise) for the jth day of the ith subject. ε_{ij} is the random error for the jth day of the ith subject. The parameters, β_k for $k = 1,2,3,4$, are the primary focus of the analysis, representing regression coefficients for age, BMI, depression, and weekend indicator respectively. Our primary goal was to accurately estimate these coefficients despite the presence of missing data. We emphasize that the particular choice of independent variables is not important to the statistical methods we developed; these variables were chosen to represent "realistic" modeling scenarios, namely mixed data-types, comprising of continuous (i.e., age, BMI) and categorical (i.e., depression, weekend status) predictors, with varying strength of association with physical

activity, as noted below.

1.4 Complete Profiles and True Model Parameters

Given our goals, the next step was to identify a "true" dataset from which we could simulate missing data patterns. We defined "complete profiles" as days with 12 or more hours of total wear time [HBK$^+$13] [HBK$^+$14]. Analytic results, in particular those of the previously discussed mixed effects regression model, derived from using only the complete profiles were considered true values.

The set of complete profiles consisted of 328 participants and 2091 days, constituting 75.7% of the total data. The number of days per participant in this set ranged from 1 to 15. We fit the linear mixed model (1.1) to the complete data set. Model parameters were estimated as $\beta_1 = -4.600$ (SE=0.833), $\beta_2 = -4.177$ (SE=1.162), $\beta_3 = -30.930$ (SE=11.672), and $\beta_4 = 8.362$ (SE=4.654). Moreover, β_1, β_2 and β_3 were significant at the 1% significance level; β_4 was significant at the 10% significance level. These regression estimates will serve as the ground-truth for testing the performance of our proposed methods.

1.5 Simulating Missing Data Patterns

The next step was to simulate "realistic" missing data patterns from complete profiles. In this section, we first describe the missing data patterns observed in our cohort, and then develop an algorithm to simulate missing data from the complete profiles. Lastly we discuss how we will apply this algorithm to test the performance of the statistical methods that will be proposed in the next section.

We observed that missing data (Figure 1.1) tended to be concentrated at the beginning and the end of a day, rather than randomly distributed throughout the day. Among the approximately

25% incomplete (i.e., with < 12 hours of recorded activity) daily profiles: 12.2% were missing 50-60% of daily records; 2.6% were missing 60-70% of daily records; 1.7% were missing 70-80% of daily records; 1.8% were missing 80-90% of daily records; 6.0% were missing 90-100% of daily records. These patterns indicate that generating simulated missing data by randomly excluding minutes of accelerometer wear throughout the day would not reflect actual missing data patterns. Hence, we designed a pairwise comparison simulation algorithm to generate realistic missing data from complete profiles. The goal was to mimic missing data patterns observed in the population, i.e., original full cohort sample. Our simulation scheme consisted of the following steps:

1 Define 5 strata of interest: days missing 50-60% data, days missing 60-70% data, days missing 70-80% data, days missing 80-90% data, and days missing 90-100% data.

2 Following the proportion of each stratum in the population data, randomly sample (without replacement) appropriate numbers of days from the set of complete profiles and assign them to a stratum. For instance, 12.2% of the complete profile set would be assigned to be missing 50-60% of their daily records.

3 For each complete profile day that is assigned to be in any of the five strata, randomly select a day of the same stratum from the original population data set and synchronize the missingness of the two days.

We illustrate the algorithm with two examples. Suppose the day sampled from the set of complete profiles had accelerometer counts recorded from 8:20 a.m. to 10:44 p.m. (Figure 1.2 top profile). Further, according to Step 2 above, assume that this complete day was assigned to be in the last stratum with 90 - 100 % missing data. Following Step 3, suppose the day of the same stratum chosen from the original data set only had accelerometer data recorded from 11:00 a.m. to 1:00 p.m.. Then our algorithm would "force" the complete profile to retain only its record from 11:00 a.m. to 1:00 p.m. (Figure 1.2 bottom profile), hence leaving the two days with the same missing

data patterns. Suppose in the next simulation, this day is assigned to be in the second stratum with 50 - 60 % missing data. Moreover, the day of the same stratum chosen from the original data set only had accelerometer data recorded from 9:00 a.m. to 8:00 p.m. Then our algorithm would "force" the complete profile to retain only its record from 9:00 a.m. to 8:00 p.m. (Figure 1.3 bottom profile), hence leaving the two days with the same missing data patterns. With this simulation algorithm in place, our study plan was straightforward. We repeated this simulation algorithm on the complete profiles 100 times, thus creating 100 datasets with missing data patterns reflected in the original population data set. We applied each of our proposed statistical methods (details below in Section 1.6) and fitted the mixed effects regression models (1.1) to each of these 100 simulated data sets. We recorded the regression coefficient estimates, as well as other key information (standard deviation, significance level, etc.) for the parameters of interest for each simulated data set. These yielded 100 estimates of these parameters, which were used to evaluate, and analytically and graphically compare, the performance of our proposed methods, in terms of mean-squared error, bias, and standard deviation.

1.6 Statistical Methods

We developed and evaluated three different analytic methods that we believed could improve the precision and accuracy of regression estimates in the presence of missing accelerometer data. We describe the rationale and statistical details of these methods below.

1.6.1 Method 1: Weighted Regression

We observed heteroskedasticity in the residuals against the proportion of missing data in a day when fitting the regression model with the raw count data (as illustrated in Figure 1.4). More specifically, the higher the amount of missing data in a day, the greater the variance of the residuals. Based on this observation, we implemented a weighted linear mixed regression model

to account for heteroskedasticity in the residuals. We considered two weighting schemes:

i variance $\propto \frac{1}{1-\tau_{ij}}$

ii variance $\propto \exp(c[i] \times \tau_{ij})$

Here τ_{ij} refers to the proportion of missing data out of 24 hours (or 1440 minutes) for subject i on day j; $c[i]$ denotes a constant c for the ith participant and is determined through a data driven built-in algorithm in the nlme package in R [PBD+15].

This method is easy to implement with slight modifications to the standard linear mixed regression model. It is also highly efficient in terms of computing time.

1.6.2 Method 2: "Imputed" Daily Sum

For the second method, we exploited the availability of multiple daily records and derived imputed estimates of total daily counts, by directly modeling the relationship between individual daily wear time and daily sum of minute-level activity counts. The rationale for this approach was to borrow information from within an individual using her multiple records, as well as, from across other participants' daily activity profiles, to obtain a more stable and accurate estimate of her own daily average activity. We developed two similar approaches:

i Fit a linear mixed regression model with daily sum of activity as the response variable and daily wear time as the predictor variable. Include a random effect for intercept but not slope, thus forcing all subjects to have the same slope $\hat{\beta}$. Use this estimated slope to impute missing data: for days with less than 12 hours (or 720 minutes) of accelerometer use, add $\hat{\beta} \times (720 - \text{wear time})$ to the observed sum of activity counts. This gives a new estimate of 12-hour total activity count for all incomplete profiles. Use these imputed activity counts per minute as the outcome in the mixed effect regression model (1.1).

ii Generalize the above method by including random effects for both the intercept and the slope for each individual. This gives an estimate $\hat{\beta}_i$ for each individual participant. Use this

subject-specific slope to impute missing data: for days with less than 12 hours of recorded data for individual i, add $\hat{\beta}_i \times (720 - \text{wear time})$ to the observed total activity count, and again fit the mixed effects regression model (1.1).

We also implemented models with a subject-specific slope term alone (i.e., with intercept equal to zero). The results were similar to method ii. above, and are not provided. Similar to the variance-weighting approach, this mixed-effects imputation method is easy to implement, using standard statistical software packages, and is highly efficient in terms of computing time.

1.6.3 Method 3: K-Nearest Neighbor

The third method imputed missing data at the minute-level even though the dependent variable of interest in our model is at the day-level (i.e., daily average activity). In order to impute missing data at the minute level, we used a K-nearest neighbor (K-NN) method, an approach that originated in the machine learning research field [HTF01]. Specifically for each incomplete profile, we would like to find the complete profiles that matched most closely with the non-missing portion of the incomplete profile. We will then use the minute level information of these complete profiles to impute the missing portion of the incomplete profile. Eventually, after imputation, all daily records would become complete profiles. To evaluate how closely two profiles match each other, we need a measure of distance between physical activity records. In this paper, we used the euclidean distance measure, but note that there is some flexibility in this choice, and other distance metrics could be used instead. The specifics of this method are described below:

- For each incomplete profile, compute its average L_2 (i.e., euclidean) distance from every complete profile. This L_2 distance is computed as the mean squares of the difference between the two profiles at every minute of the day, provided that both profiles have values at that time of the day. Then find the closest 5 neighbors, randomly select one of these neighbors, and use this neighbor's activity readings to impute missing slots.

We note that by randomly selecting one of 5-nearest neighbors, rather than the single closest neighbor, we introduced a stochastic element into the imputation, in order to avoid artificially reducing variability as is common with single imputation methods. We note also that the K-NN method can be easily extended to include additional covariates, for example day of week, weather, physical conditioning and others, so that in the imputation algorithm, nearest neighbors would be chosen to have similar values on these covariates, in addition to the activity count vectors. In summary, this minute-level imputation method is conceptually straightforward, and incorporates time of day into the imputation, and hence has the potential to be more informative and accurate than the previous methods. However, computing L_2 distance between long minute-level activity count vectors a large number of times can be computationally demanding.

1.6.4 Comparison Methods

For the sake of comparison, we also implemented other methods that are commonly used to account for accelerometer non-wear. The first method simply includes wear time (i.e., daily minutes of device wear) in the mixed effects regression model (1.1). As a second comparison, we implemented an expectationmaximization (EM) algorithm to impute average daily activity for the incomplete profiles, similar to the method by Catellier et al [CHM$^+$05]. For this EM algorithm, we considered each incomplete profile as a missing daily record and, as is recommended for missing data imputation [Sch00], we included the covariates in the regression model (1.1). We used the R package *Amelia* [HKB11] to implement the EM-algorithm. Finally, as a potentially worst-case scenario, we fit models where no adjustment was made for incomplete profiles: these models are referred to as *naïve* models in the rest of the paper.

1.7 Results

Simulation results are provided in tables and graphs. Table 1.1 gives a numerical summary comparing performances of different methods in terms of relative efficiency, bias, simulation standard deviation (sim SD), and coverage. Note that relative efficiency was calculated as the ratio of mean square error (MSE) of each method versus that of the naïve method. Lower values of relative efficiency indicated better performances. Coverage was calculated as the proportion of times the null hypothesis was rejected under the corresponding true model significance level, and the coefficient estimates had the correct sign (see Section 1.4); specifically at the 1% level for age, BMI, and depression, and 10% level for weekend. Figure 1.5 - 1.8 provide the corresponding graphical representations.

Results indicate that the naïve method, i.e., not accounting for missing data, and the conventional methods, i.e., 1) including wear time as a covariate in the model; 2) using an EM algorithm to impute average activities for the incomplete profiles [CHM$^+$05], produced similar results. These three methods had the highest values for SDs for all covariates. The two conventional methods had higher values of relative efficiency than most methods proposed in the paper. Incorporating variance weights and/or subject-level imputation with random slope (method ii) compared to naïvely ignoring missing data, reduced mean-square error by over 75%. As expected, compared to the naïve and conventional methods, incorporating variance weights improved precision, with reductions in SDs $> 50\%$, but did not reduce bias. The imputed sum methods also improved precision by $> 50\%$, with varying impacts on bias. The K-NN method improved MSE for the binary variables, but did not perform as well as the weighted regressions or the imputed sum methods. All the proposed new methods improved coverage as compared to the naïve method.

Interestingly the naïve and conventional methods exhibited lower bias compared to the other methods, especially for the continuous covariates, age and BMI. However, this lower bias

was negligible in practical terms. For example, the average bias for BMI using the conventional method of adjusting for wear time as a covariate was 0.023 vs 0.467 for the imputed sum i. method, which had the highest average bias among all the methods (Table 1.1). Converting to regression slopes, this bias translates on average to 4.177 (true model), 4.2 (adjust for wear time model) and 3.71 (imputed sum i.) lower activity counts per minute for a 1 unit higher BMI, indicating minimal differences. Of note, the best performing methods also had minimal biases, with estimated average activity count decreases of 4.106 (variance weighting i.) and 4.06 (imputed sum ii.) per unit increase in BMI. In summary, our top overall (i.e., for MSE) performers were the weighted regression with weights proportional to reciprocal of the daily wear time percentage and the imputed daily sum method ii, which incorporated a random slope in the imputation model.

1.8 Other Physical Activity Outcomes

Thus far, our analysis has focused on the total volume of activity accumulated throughout the day as the measure of physical activity. However, physical activity accumulation above or below an intensity threshold is also of interest in public health research. Hence it is important to investigate how our methods to account for missing data in accelerometry perform for other physical activity variables. In the following sections, we will examine two additional measures, namely, daily minutes of moderate-vigorous physical activity (MVPA) and daily minutes of sedentary time, both of which have emerged as important factors for health [BCM94] - [WEA+12].

1.8.1 Moderate-Vigorous Physical Activity

As in Section 1.3 our objective is to identify factors associated with MVPA in our original regression model (1.1). MVPA is measured as accelerometer counts above the threshold of 1951 counts. Variation in daily accelerometer wear time could lead to biased estimates of daily

MVPA, making it difficult to compare MVPA across days and participants. Hence we created a "standardized MVPA" variable defined as the number of minutes of MVPA in a day if the wear time during the day were 720 minutes. Thus in practice, we would estimate MVPA as $\text{MVPA} = \frac{\text{number of minutes of MVPA}}{\text{total wear time}} \times 720$. We repeated the regression analysis (Section 1.3) with this standardized MVPA as the dependent variable. As before, we fitted linear mixed models to the complete profiles, and used the estimated coefficients as the ground-truth to test the performance of our missing-data-correction methods. In the following, we will refer to this standardized MVPA variable, as simply MVPA, except when clarification is required.

Coefficient estimates for the associations between age, BMI, depression, and weekend indicator with MVPA in the complete profiles were $\beta_1 = -0.542$ (SE=0.122), $\beta_2 = -0.692$ (SE=0.170), $\beta_3 = -4.704$ (SE=1.704), and $\beta_4 = 0.101$ (SE=0.738) respectively. Moreover, β_1, β_2 and β_3 were significant at a 1% significance level; β_4 was significant at 50% significance level.

We then applied our missing data correction methods to the pseudosimulated data generated in Section 1.5. The assumptions and the methods were similar to those described in Section 1.6. The only change was in Method 2, where the response variable for imputation was now number of minutes of MVPA instead of sum of activity. We also omitted the K-NN method since it performed poorly in our previous analysis, and is computationally demanding.

Table 1.2 gives a numerical summary comparing performance of the different methods in terms of relative efficiency, bias, simulation standard deviation (sim SD), and coverage. Figures 1.9 - 1.12 provide graphical comparisons of coefficient estimates across methods. Results were generally similar to the previous results. The naïve method (i.e., not accounting for missing data) yielded unstable estimates, albeit with minimal bias. Including wear time as a covariate in the model, produced very similar results to the naïve method. Using an EM algorithm to impute MVPA for the incomplete profiles had smaller MSE than the naïve or conventional adjustment method, but was generally more biased. Imputation daily sum method i. with random intercepts (and fixed slope) had the highest bias for three of the predictors, but had lower MSE than the

conventional methods. Subject-level imputation with random slope (method ii) compared to ignoring missing data, generally reduced mean-square error by $> 30\%$, and improved precision and coverage. The weighted regressions performed the best with $> 60\%$ reductions in MSE and simSD, and negligible bias compared to the naïve method; the top performer was weighted regression with weights proportional to reciprocal of the daily wear time percentage.

1.8.2 Sedentary Time

In this section we examined how our correction methods performed if sedentary time was the outcome. Sedentary time is measured as accelerometer counts below the threshold of 100 counts. Again, our estimate of daily sedentary time took into account the total wear time during that day. Thus "standardized sedentary time" was defined as the number of minutes of sedentary time in a day if the wear time during the day were 720 minutes, i.e. Sedentary Time $= \frac{\text{number of minutes of sedentary time}}{\text{total wear time}} \times 720$. We repeated the regression analysis (Section 1.3) with this standardized sedentary time as the dependent variable. As before, we fitted linear mixed models to the complete profiles, and used the estimated coefficients as the ground-truth to test the performance of our missing-data-correction methods. In the following, we will refer to this standardized sedentary time variable, as simply sedentary time, except when clarification is required.

Coefficient estimates for the associations between age, BMI, depression, and weekend indicator with sedentary time in the complete profiles were $\beta_1 = 1.979$ (SE=0.550), $\beta_2 = 1.793$ (SE=0.767), $\beta_3 = 7.169$ (SE=7.704), and $\beta_4 = -8.852$ (SE=2.953) respectively. Moreover, β_1 and β_4 were significant at 1% significance level; β_2 was significant at 5% significance level; β_3 was significant at the 50% significance level.

We then applied our missing data correction methods to the pseudosimulated data generated in Section 1.5. The assumptions and the methods were similar to those described in Section 1.6. The only change was in Method 2, where the response variable for imputation was now

number of minutes of sedentary time instead of sum of activity. We again omitted the K-NN method since it performed poorly in our previous analysis, and is computationally demanding.

Table 1.3 gives a numerical summary comparing performance of the different methods in terms of relative efficiency, bias, simulation standard deviation (sim SD), and coverage. Figures 1.13 - 1.16 provide graphical comparisons of coefficient estimates across methods. The naïve method, i.e., not accounting for missing data, and the conventional methods, i.e., 1) including wear time as a covariate in the model and 2) using EM algorithm to impute sedentary time for incomplete profiles, produced similar, if not worse, results. These methods exhibited the smallest bias, especially for the continuous covariates, but large SDs for all the covariates, resulting in large mean-squared errors. Imputed daily sum method i. with random intercepts generally had the highest bias. Each of the proposed new methods, i.e., incorporating variance weights or subject-level imputation, improved precision (simSD) by $> 40\%$, and improved coverage by $> 25\%$. Our top overall performers in terms of mean-squared error were the weighted regression with weights proportional to reciprocal of the daily wear time percentage and the imputed daily sum method ii, which incorporated a random slope in the imputation model; these methods, compared to the conventional methods, reduced MSE by $>75\%$ for all the covariates.

1.9 Theoretical Underpinning: a Poisson Framework

The variance weighting method with weights inversely proportional to daily wear time percentage generally had the least error across all our analyses. We can justify this superior performance under a plausible Poisson generating mechanism for the accelerometer data. In particular, suppose that activity count at each minute follows a Poisson distribution with intensity parameter λ. We further assume that the proportion of daily wear time for subject i converges to a constant, denoted as $\theta_i = \theta$. Now suppose that on a given day, we observe $m_i = m$ minutes of activity for subject i (note: to reduce complexity of notation, we suppress the index i when

referring to θ and m). We naïvely use the mean of these m Poisson activity readings to estimate the average activity for this day for subject i. Theoretically, under an iid assumption, this estimator has mean λ and variance λ/m, which by our assumption will converge to $\lambda/(\theta \times T)$, where T denotes the ideal total number of minutes of device wear time in a day. Therefore, it is reasonable to set the variance of our regression model to be proportional to $1/\theta$. Although this is a simplified model, we believe that using this Poisson framework provides a useful theoretical heuristic to explain the superior performance of the variance-weighted model.

1.10 Conclusions

Accurate measurement of physical activity is a critical factor for designing and implementing interventions aimed at modifying this important behavior. Accelerometers provide rich and objective data on individual-level physical activity patterns throughout the day, and hence have emerged as an important tool in physical activity research. Despite their many advantages, analysis of accelerometer data presents many challenges [SZC12].

In this paper, we focused on one specific challenge, namely missing data due to non-wear. Using a large cohort, we characterized patterns of accelerometer non-wear during the day, which is a critical first step to developing methods to correct for missing data. We observed that missing data patterns tended to occur at the beginning and end of the day. This observation is important for simulating realistic missing data patterns, i.e., randomly distributed non-wear throughout the day is atypical, and assuming a completely-at-random missingness pattern would not mimic what occurs in practice. Next, we used a pseudosimulation approach to develop and compare three new statistical approaches to correct for missing accelerometer data in a regression model setting with physical activity is the dependent variable of interest, such as in an intervention trial. For the sake of comparison, we also implemented two existing methods commonly used to account for device non-wear: including wear time as a covariate, and an EM-type imputation algorithm [CHM$^+$05].

Our results indicated that ignoring missing data had a major impact on precision (i.e., high standard errors of regression parameters). The conventional methods of including wear time as a covariate or an EM-type imputation did not improve precision. Among the proposed methods, (a) variance weighting by the inverse of daily wear time proportions and (b) imputing activity using subject-specific intercepts and slopes, improved precision the most, albeit with a small increase in bias for some covariates. However, the gain in precision far outweighed the increased bias, which was negligible in practical terms. Improving precision of regression estimates has major implications for study design: large variability requires larger sample-sizes to ensure adequate power. Equivalently, ignoring non-wear or controlling for wear time, may result in failure to reject a null hypothesis. Of note, our variance weighting method has a theoretical basis under a Poisson framework, while the imputation method exploits the hierarchical nature of the data, namely day-level information nested within subjects.

A few limitations need to be noted. Our cohort comprised of postmenopausal breast cancer survivors, which could limit generalizability. It is possible that a different sample may exhibit a different missing data pattern. However, other studies ([CHM$^+$05],[SRJL$^+$15]) have noted concentrated missing data patterns, similar to what we observed. Second, our sample used a wake-time protocol for hip-worn accelerometers. Wrist-worn and small hip-worn accelerometers, which can be worn for 24-hours, are gaining in popularity and may have fewer problems with device non-wear, since participants are more likely to wear the devices for longer periods. But 24-hour protocols do not result in 24 hours of wear, and variation in wear time may, in fact, increase with longer wear periods, making it even more important to consider statistical adjustments for non-wear when analyzing these data. Furthermore, intensity cutpoints for wrist accelerometers have not been standardized, and many different algorithms have been proposed for classifying activity level ([EKG$^+$15], [HNS$^+$16], [SSB$^+$00]). Besides, hip-worn devices have been used to measure physical activity in several large existing cohorts comprising over 50,000 participants [LS14], including subsets of the NHANES database, and the OPACH substudy of the Women's

Health Initiative. These unique and well-characterized databases include information on a variety of cardiovascular, cancer, psychosocial and other outcomes, resulting in unique data sources for examining physical activity-disease associations. Hence statistical methods for missing data for these hip devices are still highly relevant in public health research. Importantly, our proposed statistical approaches are generalizable, and we do not expect the choice of cohort or device to have a major impact on how these new statistical methods will perform. Thirdly, our analysis focused on physical activity as a dependent variable, and examined its correlates. We did not investigate if our methods would also correct biases in models where physical activity is an independent variable. Although the mixed-effects and K-NN imputation methods are easily applicable to the latter case, such analyses are fundamentally different from our setting, and we leave their investigation to a future project. Finally, we note that statistical methods using functional data analysis ([MAC+06b], [XHS+15]) to model minute-level accelerometer data are being increasingly proposed. These methods implement minute-level interpolation or imputation via likelihood or Bayesian methods to account for non-wear. While these minute-level methods hold promise, they are complex and computationally challenging. Of note, and perhaps surprisingly, in our simulations, the EM-algorithm, a standard method for missing data imputation, and K-NN, which considers time-of-day in the imputation algorithm, both were outperformed by the simpler variance weighting and imputed sums methods. A possible reason for this is that when the unit of analysis is day-level activity, these more complex methods may actually introduce more variance. Thus, the methods we propose are easy to implement using standard software and hence will be accessible to the applied statistician or epidemiologist.

In summary, we introduce two new methods to account for non-wear in accelerometer-based physical activity research. We considered three measures of physical activity: total volume summarized as daily accelerometer counts per minute, daily minutes of moderate-vigorous physical activity, and daily minutes of sedentary time. Our results were consistent across the three activity measures, and indicated that implementing variance-weighting or imputation using

subject-specific parameters could vastly reduce variability in parameter estimates, thus improving study power. There is a bias-variance trade-off whereby the proposed methods could lead to increased bias, but as noted these biases were negligible and of little practical importance. We anticipate that these easy-to-implement correction methods will be useful in physical activity research.

1.11 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in "Statistical approaches to account for missing values in accelerometer data: Applications to modeling physical activity", *Statistical Methods in Medical Research*, Vol 27, Issue 4, pp.1168-1186. Selene Yue Xu, Sandahl Nelson, Jacqueline Kerr, Suneeta Godbole, Ruth Patterson, Gina Merchant, Ian Abramson, John Staudenmayer, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Table 1.1: Comparison of Methods for Total Physical Activity. Relative efficiency: ratio of mean-squared error between the current method and the naÏve model; Coverage: percentage of times null is rejected under the corresponding true model significance level, and the coefficient estimates have the correct sign; "NaÏve model" disregarded missing data. "Adjust for wear time" model included wear time as a covariate. "EM imputation" model used EM algorithm to impute average activities for the incomplete profiles.

		Relative Efficiency	Bias	Sim SD	Coverage
NaÏve Model	age	1.000	-0.053	0.317	1.00
	BMI	1.000	-0.010	0.427	0.98
	depression	1.000	0.855	4.985	0.30
	weekend	1.000	0.670	5.338	0.30
Adjust for Wear Time	age	1.456	-0.027	0.387	1.00
	BMI	1.074	0.023	0.442	0.97
	depression	0.993	0.780	4.979	0.31
	weekend	1.040	1.230	5.347	0.30
EM Imputation	age	0.824	0.036	0.289	1.00
	BMI	1.356	0.198	0.456	1.00
	depression	0.982	1.150	4.879	0.88
	weekend	0.784	-0.022	4.764	0.39
Weighted Regression i	age	0.208	-0.082	0.121	1.00
	BMI	0.161	0.071	0.156	1.00
	depression	0.123	0.769	1.601	0.49
	weekend	0.178	1.060	2.006	0.77
Weighted Regression ii	age	0.238	-0.089	0.129	1.00
	BMI	0.208	0.098	0.168	1.00
	depression	0.231	1.625	1.805	0.34
	weekend	0.236	1.191	2.326	0.68
Imputed Sum i	age	1.425	0.360	0.131	1.00
	BMI	1.365	0.467	0.174	1.00
	depression	0.558	3.290	1.856	0.53
	weekend	0.176	0.156	2.250	0.62
Imputed Sum ii	age	0.144	0.007	0.122	1.00
	BMI	0.207	0.116	0.155	1.00
	depression	0.128	0.820	1.615	0.53
	weekend	0.124	0.033	1.893	0.76
K-NN	age	0.975	0.272	0.163	1.00
	BMI	1.473	0.450	0.256	1.00
	depression	0.596	3.136	2.328	0.34
	weekend	0.297	-0.056	2.932	0.52

Table 1.2: Comparison of Methods for Moderate-Vigorous Physical Activity. Relative efficiency: ratio of mean-squared error between the current method and the naÏve model; Coverage: percentage of times null is rejected under the corresponding true model significance level, and the coefficient estimates have the correct sign; "NaÏve model" disregarded missing data. "Adjust for wear time" model included wear time as a covariate. "EM imputation" model used EM algorithm to impute MVPA for the incomplete profiles.

		Relative Efficiency	Bias	Sim SD	Coverage
NaÏve Model	age	1.000	-0.014	0.054	1.00
	BMI	1.000	0.009	0.088	0.96
	depression	1.000	-0.080	0.844	0.47
	weekend	1.000	0.100	1.095	0.26
Adjust for Wear Time	age	1.073	-0.014	0.057	0.98
	BMI	0.825	-0.013	0.079	0.96
	depression	0.930	0.042	0.817	0.40
	weekend	0.955	0.077	1.072	0.29
EM Imputation	age	0.806	0.023	0.045	1.00
	BMI	0.785	0.054	0.057	1.00
	depression	0.676	0.185	0.672	0.94
	weekend	0.417	-0.098	0.703	0.27
Weighted Regression i	age	0.161	-0.012	0.019	1.00
	BMI	0.085	0.008	0.024	1.00
	depression	0.089	0.015	0.253	0.87
	weekend	0.104	0.153	0.320	0.58
Weighted Regression ii	age	0.232	-0.017	0.021	1.00
	BMI	0.120	0.016	0.026	1.00
	depression	0.136	0.122	0.288	0.70
	weekend	0.232	0.412	0.331	0.48
Imputed Sum i	age	0.760	0.044	0.021	1.00
	BMI	0.774	0.074	0.025	1.00
	depression	0.341	0.416	0.269	0.84
	weekend	0.104	0.037	0.352	0.52
Imputed Sum ii	age	1.049	0.007	0.057	0.99
	BMI	0.721	0.024	0.071	0.99
	depression	0.398	0.115	0.523	0.87
	weekend	0.075	0.030	0.299	0.52

Table 1.3: Comparison of Methods for Sedentary Time. Relative efficiency: ratio of mean-squared error between the current method and the naÏve model; Coverage: percentage of times null is rejected under the corresponding true model significance level, and the coefficient estimates have the correct sign; "NaÏve model" disregarded missing data. "Adjust for wear time" model included wear time as a covariate. "EM imputation" model used EM algorithm to impute MVPA for the incomplete profiles.

		Relative Efficiency	Bias	Sim SD	Coverage
NaÏve Model	age	1.000	0.003	0.210	1.00
	BMI	1.000	-0.032	0.259	0.73
	depression	1.000	-0.646	3.102	0.61
	weekend	1.000	-0.418	2.903	0.29
Adjust for Wear Time	age	0.988	0.003	0.209	1.00
	BMI	0.996	-0.011	0.260	0.78
	depression	1.008	-0.633	3.118	0.61
	weekend	0.975	-0.227	2.886	0.26
EM Imputation	age	1.108	0.015	0.220	1.00
	BMI	1.776	-0.120	0.327	0.89
	depression	1.189	-0.308	3.442	0.73
	weekend	1.287	0.592	3.274	0.39
Weighted Regression i	age	0.163	0.004	0.084	1.00
	BMI	0.202	-0.041	0.110	0.98
	depression	0.128	-0.455	1.036	0.91
	weekend	0.194	-0.231	1.272	0.81
Weighted Regression ii	age	0.203	-0.004	0.094	1.00
	BMI	0.212	-0.046	0.111	0.98
	depression	0.157	-0.723	1.026	0.85
	weekend	0.224	0.405	1.326	0.51
Imputed Sum i	age	0.938	-0.185	0.083	1.00
	BMI	0.841	-0.195	0.139	0.92
	depression	0.243	-0.977	1.218	0.88
	weekend	0.260	0.229	1.479	0.72
Imputed Sum ii	age	0.174	-0.027	0.083	1.00
	BMI	0.231	-0.056	0.112	0.97
	depression	0.120	-0.357	1.039	0.93
	weekend	0.169	0.378	1.143	0.84

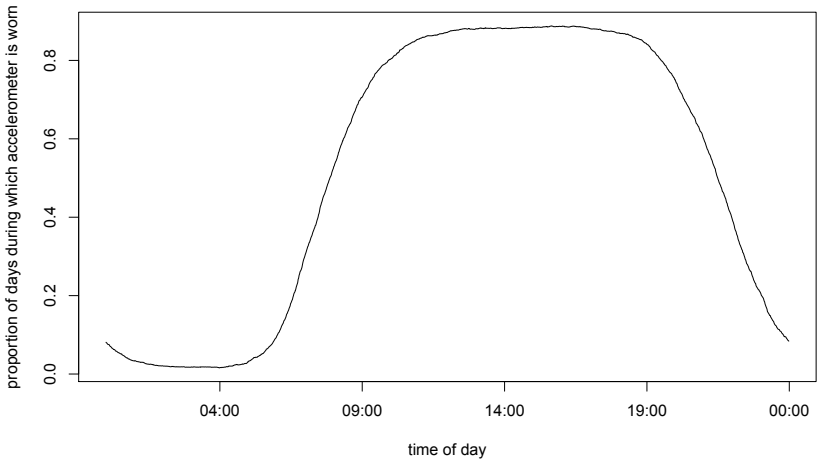

Figure 1.1: Proportion of days during which accelerometer was worn at each time point of the day.

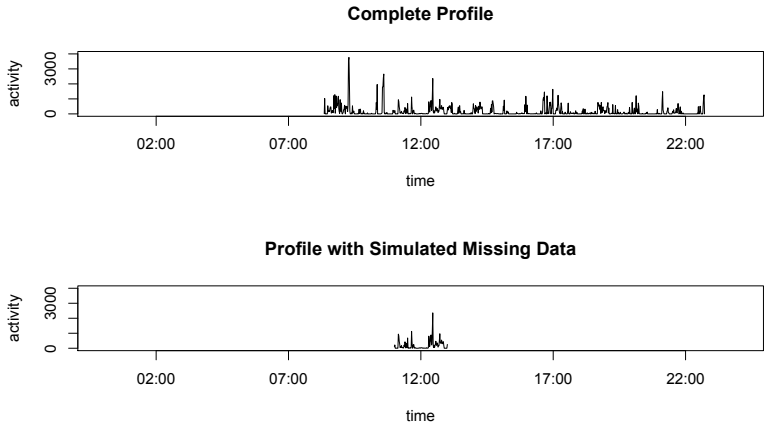

Figure 1.2: First Example Illustrating the Missing Data Simulation Algorithm

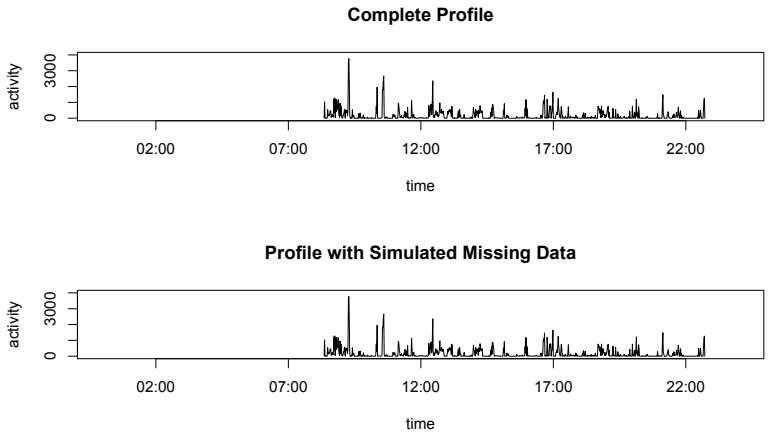

Figure 1.3: Second Example Illustrating the Missing Data Simulation Algorithm

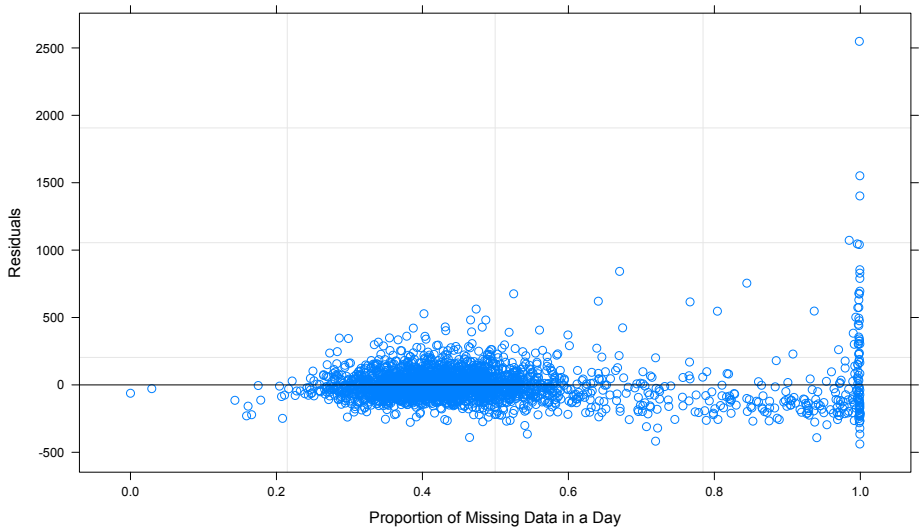

Figure 1.4: Heteroskedasticity in the Residuals against the Proportion of Missing Data in a Day

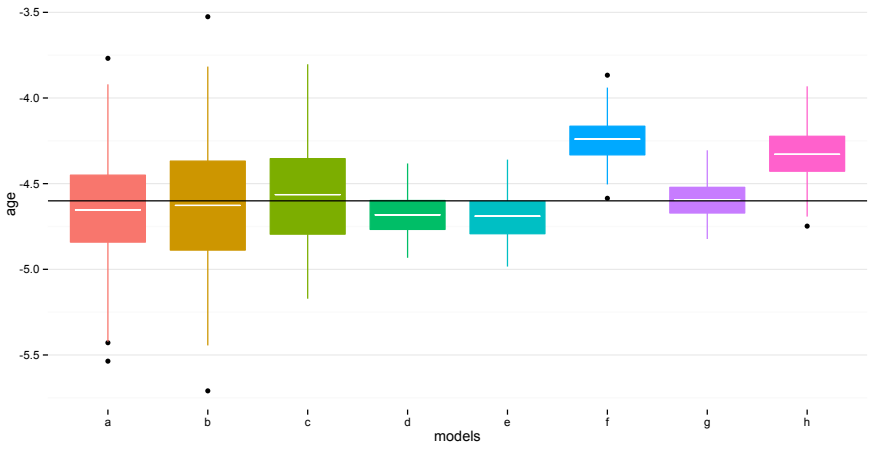

Figure 1.5: Compare performances of different methods in estimating the age coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii; h) K-NN.

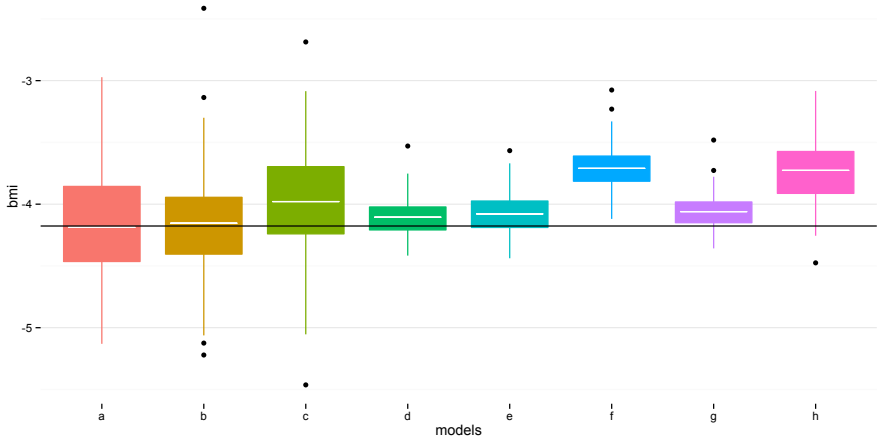

Figure 1.6: Compare performances of different methods in estimating the BMI coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii; h) K-NN.

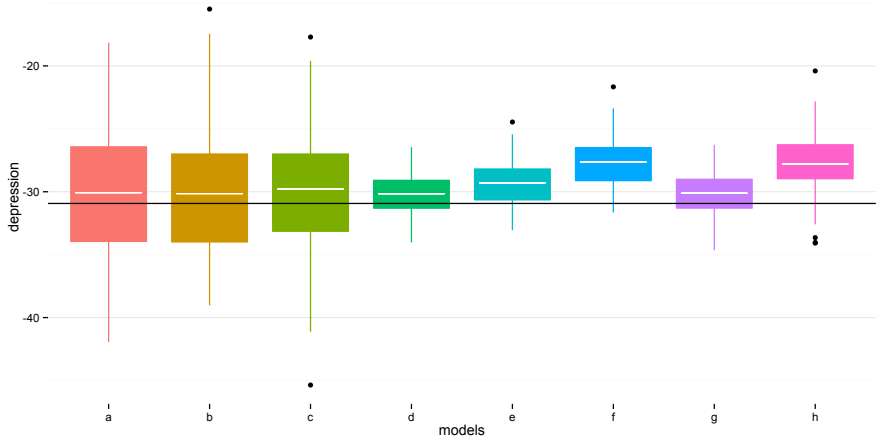

Figure 1.7: Compare performances of different methods in estimating the depression indicator coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii; h) K-NN.

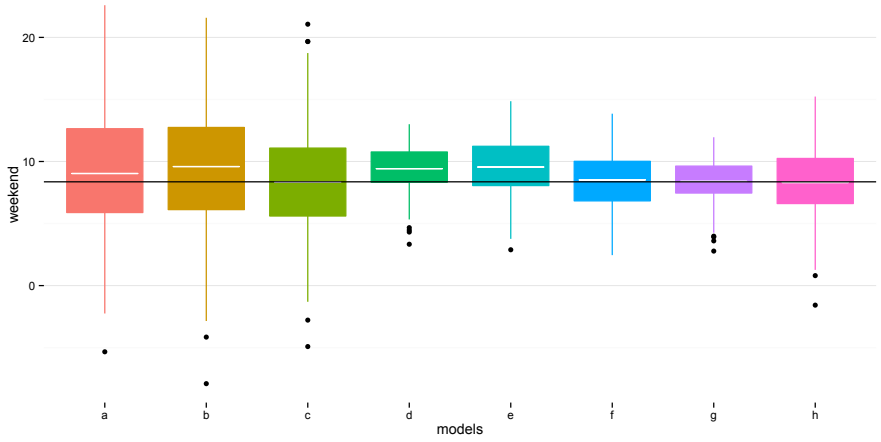

Figure 1.8: Compare performances of different methods in estimating the weekend indicator coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii; h) K-NN.

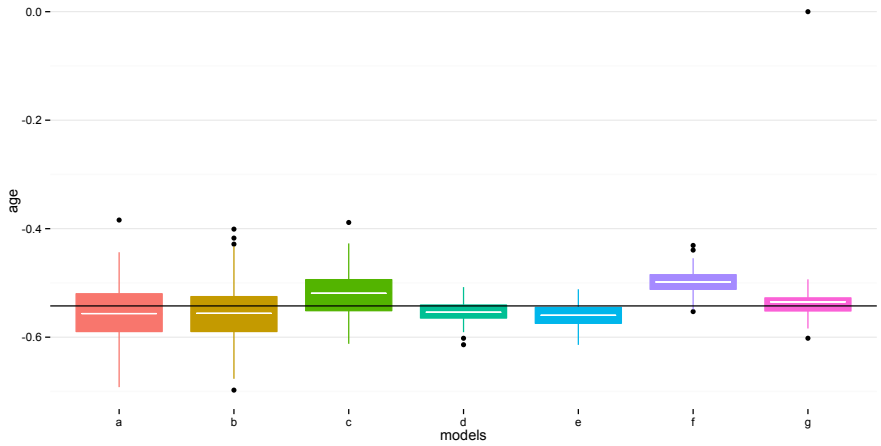

Figure 1.9: Compare performances of different methods in estimating the age coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

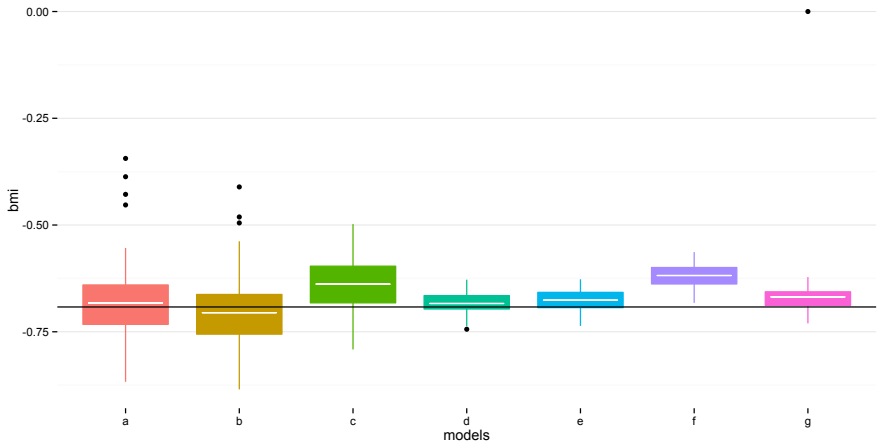

Figure 1.10: Compare performances of different methods in estimating the BMI coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

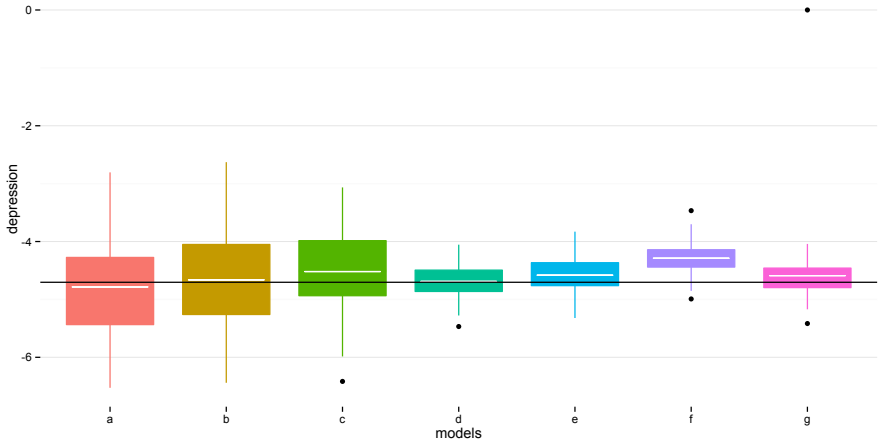

Figure 1.11: Compare performances of different methods in estimating the depression indicator coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

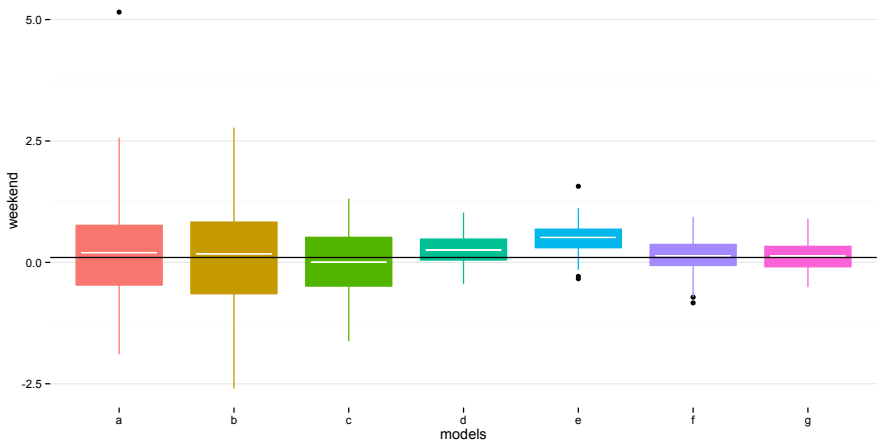

Figure 1.12: Compare performances of different methods in estimating the weekend indicator coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

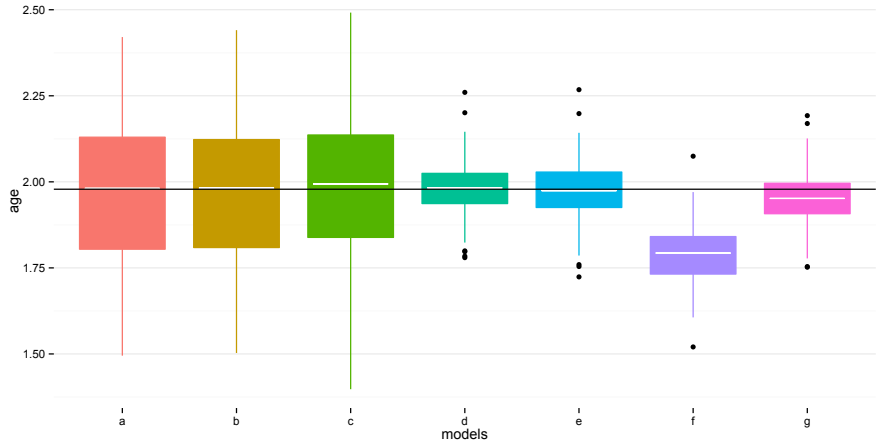

Figure 1.13: Compare performances of different methods in estimating the age coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

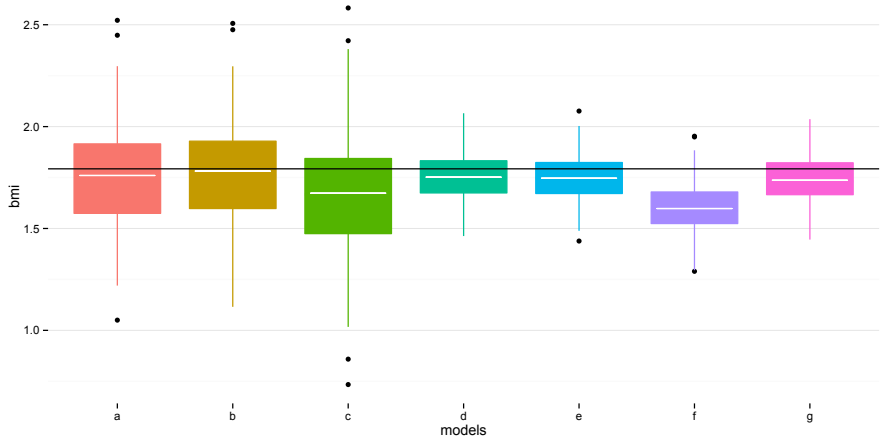

Figure 1.14: Compare performances of different methods in estimating the BMI coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

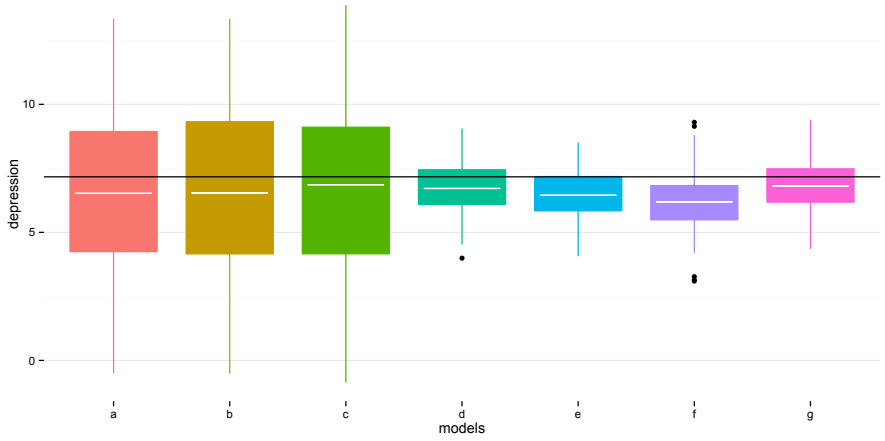

Figure 1.15: Compare performances of different methods in estimating the depression indicator coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

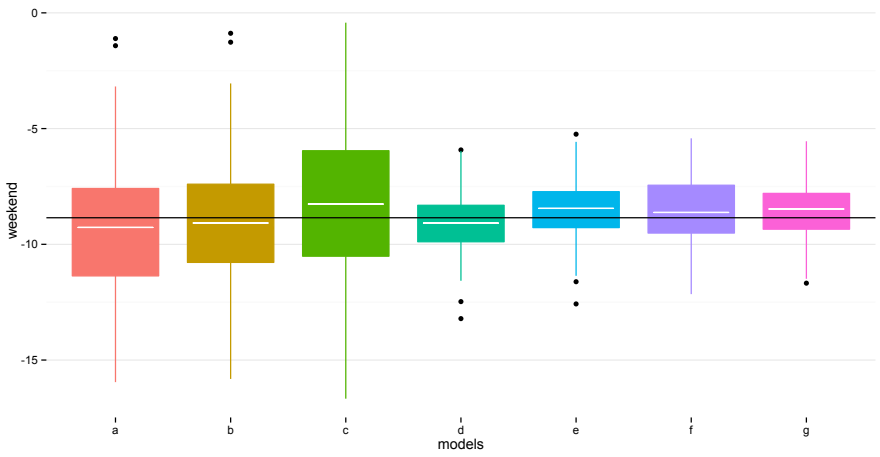

Figure 1.16: Compare performances of different methods in estimating the weekend indicator coefficient. a) naïve model; b) adjust for wear time; c) EM imputation; d) weighted regression i; e) weighted regression ii; f) imputed sum i; g) imputed sum ii.

Chapter 2

Bayesian Network Project - Sleep Study

2.1 Introduction

Cancer patients report a cluster of fatigue, sleep, and mood problems before and during adjuvant cancer treatment [SBB+14] [WH15]. Cancer-related fatigue, one of the most distressing symptoms, is characterized by extreme tiredness. Cancer patients also report sleep problems, e.g., difficulties with falling and staying asleep, before, during, and after chemotherapy [AILM+06] [BBB+10] [Bow08]. Insomnia and sleep problems have been associated with fatigue [AILM+06] [BM08] [WBH09], depression, and decreased quality of life during treatment [AILM+06] [BBBC06] [AILR+14] [O'D04]. Depression is common [ALC+15] [SML+09] with 40-82% of patients undergoing chemotherapy reporting clinically significant depressive symptoms [Mas04]. Importantly, our lab has shown that in women with breast cancer, all symptoms within the symptom cluster, i.e., fatigue, sleep complaints and depressive symptoms, whether or not present before chemotherapy, worsen during cancer treatment [AILM+06] [LFN+09].

Chemotherapy-related cognitive problems are frequently reported by cancer survivors [ARS+16] [CMT+13] [HAS+12]. In a recent meta-analysis, 15 (12 in breast cancer) of 17 studies observed objective cognitive decline in patients treated with chemotherapy [DRH+17].

These problems can last for a few weeks, months or even years after completion of chemotherapy [GKC+13]. Imaging studies have found alterations in cerebral activity in cancer patients before and after chemotherapy [MLF+17]. It is likely that the cognitive impairment seen in cancer patients might, at least in part, be related to fatigue, sleep problems and depression [VRW+09].

Unraveling interrelationships between these symptoms and their impact on cognition poses computational challenges. Dimension reduction methods such as principal component analysis, or clustering techniques are useful for deriving summary statements regarding association between the symptom cluster as a whole and cognitive symptoms in cancer patients. However, these methods use weighted combinations of suitably normalized symptom cluster variables, rendering results difficult to interpret for individual symptoms. While standard regression modeling cannot disentangle these complex associations, moderated regression methods could be used in this context [Fra15a] [Fra15b]. Yet, alternative novel methods are needed to assess robustness of findings.

Bayesian graphical networks are a powerful approach for examining multivariate relationships. Bayesian networks provide algorithms for discovering and analyzing structure with intuitive graphs for visualizing interrelationships among sets of variables. The initial development of Bayesian networks arose in computer science and artificial intelligence [Pea14], and since then it has become widely used in genomics, and medical applications [TB07] [VPR+07].

Bayesian networks have not been widely used in cancer symptoms research. Herein, we illustrate how to apply Bayesian networks to examine associations amongst symptoms related to chemotherapy treatment, and their role in cognitive dysfunction, and quality of life. Our goal is to demonstrate how this powerful computational method can be used to explore complex interrelationships between variables, and possibly guide design of intervention studies.

2.2 Brief description of Bayes Network Methodology

A Bayesian network is a statistical model that represents multivariate relationships between sets of variables via a graph. Nodes on the graph depict random variables, while edges represent dependencies between variables. Each node has an associated probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. The presence of an edge or path between two variables indicates a non-zero (partial) correlation between the two variables [Pea14]. Specifically, a Bayesian network analysis derives observed probabilistic dependencies and (conditional) independencies between sets of variables. Under certain technical assumptions [Pea14] [NSL13] these observed probabilistic relationships conform to what could have been observed from a hypothetical causal network with the same structure under a controlled experiment, where each variable is manipulated while holding others constant. Thus a Bayesian network can generate hypotheses that can be tested in future studies.

Fitting a Bayesian network requires learning its structure (i.e., which nodes in the graph are connected) and parameters (i.e., estimation of conditional probabilities). Specifically, let X comprise the set of variables X_i (e.g., X_1 = fatigue, X_2 = sleep quality, X_3 = mood, X_4 = cognition) and M be a Bayesian network on X, i.e., a graph of edges between variables in X, as in Figure 2.1. The model M encodes conditional independencies that imply a factoring of the joint probability distribution $p(X)$ of X:

$$Pr(X|M) = \prod Pr(X_i|pa(X_i)) \tag{2.1}$$

where $pa(X_i)$ denotes variables ("parents") in X with directed edges (i.e., arrows) leading to X_i. To learn the structure of the graph M (e.g., links, or edges, between fatigue, sleep, mood, and cognition in our application), efficient constraint-based, search-score, and hybrid algorithms can be implemented [NSL13]. For a given graph M, parameters β of $p(X)$, can be estimated by regression methods, using multivariate Gaussian (after appropriate transformation if needed) or

nonparametric distributions for continuous, and multinomial distributions for categorical variables [NSL13].

2.3 Methods

2.3.1 Study Sample and Measures

We conducted a secondary analysis of an existing database of a completed NIH-funded study (2004-2010) on the relationship between sleep, fatigue, mood, and cognition in breast cancer patients (PI Ancoli-Israel). Details on the study design and protocol have been previously published [AILR$^+$14]. The study was approved by the UCSD Human Research Protections Committee (protocol #s 080120 and 120187) and the UCSD Moores Cancer Centers Protocol Review and Monitoring Committee. We briefly describe the study sample and measures pertinent to our study.

Study Sample

The study recruited 74 newly diagnosed stage I-III breast cancer patients (mean age=51.8y) who were scheduled to receive chemotherapy, and followed them for one year. Data were collected at three time points: before the start of chemotherapy (baseline; BL), at the end of cycle 4 chemotherapy (C4) and one year after the start of chemotherapy (Y1). In order to reduce confounding by other medical conditions or medications, the study excluded pregnant women, patients with significant anemia, patients currently receiving radiotherapy or bone marrow transplants or treatment for sleep apnea or periodic limb movements in sleep. Also, patients with current diagnosis of major depression, anxiety or psychotic disorder, and patients using medications known to influence sleep for three months prior to enrollment, were excluded.

Participant characteristics are provided in Table 2.1. Briefly, 88% were white, and 51% were college graduates, and close to 70% had Stage I or II cancers. The median (25th, 75th

percentile) interval between the pre-chemotherapy and end-of-chemotherapy assessments was 79 (64, 84) days, and between the end-of-chemotherapy and the year 1 measures was 341 (317, 409) days.

Symptom Cluster and Psychosocial Functioning Assessment

Symptoms were assessed via validated questionnaires. At BL, C4 and Y1, patients self-reported sleep quality on the Pittsburgh Sleep Quality Index (PSQI) [BRM+89]; fatigue on the Multidimensional Fatigue Symptom Inventory-Short Form (MFSI-SF) [SMHJ98]; mood on the Center of Epidemiological Studies-Depression (CES-D) scale [Rad77]; quality of life (QoL) on the FACT-B scale for breast cancer patients and functional-outcomes-of-sleepiness (FOS) on the functional-outcomes-of-sleepiness questionnaire (FOSQ) [WLE+97]. FOS assesses how sleepiness impacts daily functioning.

Neurophyschological (NP) Testing

Cognitive function was assessed with an objective NP test battery, which targeted a number of specific cognitive abilities associated with chemotherapy related impairment, including episodic learning/memory, attention/working memory, executive functions, and psychomotor/processing speed. Specific component tests in this battery included Digit Span, Digit Symbol, and Symbol Search subtests from the Wechsler Adult Intelligence Scale-Third edition (WAIS-III) [Wec14]; Trail Making Tests A and B [RW93]; Hopkins Verbal Learning Test-Revised [BSGB98]; Wisconsin Card Sorting Task - 64 card version (conceptual level of responses) [KTIH00]; Stroop Color-Word Interference test (interference trial) [GF02]; and the Letter and Category [GSE+99] (animals) Fluency test (total words generated).

A summary measure of cognitive ability, a NP composite score, was computed as follows: each component raw test score was converted to a z-score by subtracting the baseline mean and dividing by the standard deviation. Z-scores were coded so that higher scores represented better

functioning, and the composite score was defined as the mean of z-scores over the entire battery. As supplementary material, mean (SD) of the individual test scores at each study time-point are provided in Table 2.4.

Previous analyses of longitudinal symptoms and cognition

We have previously shown [AILR$^+$14] that compared to healthy controls, breast cancer patients had worse sleep quality (PSQI), more fatigue (MFSI), worse mood (CESD), worse functional-outcomes-of-sleepiness (FOSQ) and Quality of life prior to chemotherapy. Also, these factors worsened for the patients during chemotherapy, compared to controls. By Y1, symptoms in the patients were not different to their baseline values, but were still worse compared to controls.

In summary, our prior analyses examined each symptom individually and demonstrated that, on average, sleep quality, fatigue, mood, and QoL worsened in breast cancer patients during chemotherapy. However, this previous work did not evaluate how these different symptoms influenced each other and cognition over time. The Bayesian network analyses proposed below aims to address this latter question.

2.3.2 Statistical Methods

We calculated summary statistics of demographic factors, as well as, mean (SD) at BL, C4 and Y1 for the symptom cluster (sleep quality, fatigue, mood), QoL, and cognition.

We fit a Bayesian network to examine multivariate relationships between the symptom cluster, quality of life and cognition, before, during and after chemotherapy. We also included demographic variables (age, college educated [yes vs no]) in the network. Our network included pre-chemotherapy (BL), post-chemotherapy (C4), and year 1 follow-up measures (Y1), and examined temporal and cross-section relationships among variables. In particular, measures at time t were allowed to have directed edges to measures at time (t+1), but not vice versa.

We used a score-based hill-climbing algorithm to infer network structure, and applied bootstrap resampling to learn a set of 500 network structures. We then averaged these networks in an effort to reduce the impact of locally optimal (but globally suboptimal) networks on learning and inference. The averaged network is a more robust model with better predictive performance than choosing a single, high-scoring network [NSL13]. To quantify stability of inferred edges, we computed arc strength and direction strength. Arc strength was calculated as the frequency of an edge occurring between two variables across the 500 bootstrapped network structures; similarly, directional strength was assessed as the frequency of the observed direction re-occurring in the set of learned network structures. We inferred conditional independencies between variables via the theory of Markov blankets of networks [NSL13]. We applied Bayesian information criteria (BIC) and posterior model probabilities to compare candidate networks [Was00]. Lower BIC scores indicate better fit; score differences > 5 (respectively between 2.2 and 5) between two models strongly (respectively moderately) favor the lower-scoring model; differences < 2.2 indicate similar fit for both models. Models were fitted using the R package bnlearn [NSL13].

2.4 Results

Longitudinal scores (Table 2.2) indicate that, as noted previously [AILR$^+$14], symptoms and QoL worsened during chemotherapy (BL-C4) on average, but were generally comparable to BL levels by Y1. Cognitive performance did not change significantly during chemotherapy, but was significantly higher at year 1 compared to BL.

2.4.1 Bayesian Network Results

Decomposition of probability distribution

Using the derived network, we decomposed the joint probability distribution of all 20 variables (2 demographic, and 6 symptoms/outcomes at 3 time-points) as a product of conditional

distributions. Specifically, letting X represent the vector of all the variables, we have the following factorization. The suffix 0 is BL, 1 is C4 and 2 is Y1.

$$\begin{aligned} P(X) = P[Mood0] \times P[Mood2] \times P[Age] \times P[Education] \times P[Cognition0|Age] \\ \times P[Fatigue0|Mood0] \times P[QoL2|Mood2] \times P[QoL0|Fatigue0] \times P[FOS0|Fatigue0] \\ \times P[Mood1|QoL0] \times P[Sleep0|QoL0] \times P[Fatigue1|Mood1] \\ \times P[QoL1|Mood1 + Fatigue1 + QoL0] \times P[Sleep1|Fatigue1 + Sleep0] \\ \times P[FOS1|Fatigue1 + QoL0 + FOS0] \times P[Cognition1|Cognition0 + Sleep1] \\ \times P[Fatigue2|Fatigue1 + QoL2 + Sleep1] \times P[FOS2|QoL2 + FOS1] \\ \times P[Cognition2|Cognition0 + Cognition1] \times P[Sleep2|Mood2 + Sleep0 + FOS2] \quad (2.2) \end{aligned}$$

$P(A|B)$ denotes the conditional distribution, i.e., probability of a variable A, given that we know the value of variable B. Thus, the above decomposition converts the complex model comprising 20 variables into simpler "local" components, and highlights subsets of factors that directly influence each variable. In fact, in our network (Figure 2.1), the maximum number of parents, i.e., directed edges pointing to any variable, is 3 (e.g., Fatigue at Y1 has parents Fatigue and Sleep at C4, and QoL at Y1), thus substantially fewer than the maximum of 19 possible directed edges. Below, we highlight key findings, and describe how to infer (in)dependencies between variables.

Cognitive functioning and symptoms

The network elicits local structure, so that we can identify parents, namely variables that directly influence any given factor. For e.g., age was the parent of BL cognition, whereas BL cognition and C4 sleep quality were parents of C4 cognition. The bootstrapped arc strength for

age on BL cognition was 0.86 indicating that age was reproducibly associated with cognition. The regression estimate (Table 2.3) was negative for age indicating that younger age was associated with better cognition at BL.

At C4, cognition was (directly) positively influenced by BL cognition (as might be expected) but negatively influenced by C4 sleep score, indicating that worse sleep quality at the end of chemotherapy was associated with worse cognition. Interestingly, via the Markov property, we infer that after accounting for cognition and C4 sleep, C4 cognitive function was independent of all other variables. We also note that although not directly linked, C4 fatigue affected C4 cognition through C4 sleep quality. Moreover, C4 depression indirectly affected cognition through a direct effect on C4 fatigue, and corresponding downstream effects on C4 sleep quality.

At the one year follow-up, cognition was directly influenced by both BL cognition and C4 cognition, with, as expected, positive regression coefficients for both variables, indicating that higher BL and C4 cognition scores were associated with better Y1 cognition. Interestingly, no symptoms directly influenced Y1 cognition.

Symptom clusters, quality of life (QoL) and functional outcomes of sleepiness (FOS)

Focusing on BL Quality of life and functional outcomes of sleepiness, BL fatigue was the only parent of BL QoL (arc-strength = 0.93) and BL FOS (arc-strength = 0.70), with negative regression estimates, indicating that less fatigue was associated with better QoL and FOS (Table 2.3).

At C4, there were several parents for each of QoL and FOS. C4 mood and fatigue, and BL QoL were all parents of C4 QoL, with C4 mood exhibiting the most consistent effects (arc strength = 0.94, arc direction = 0.89). Also, via the Markov property we can infer that after conditioning on BL QoL, C4 mood and C4 fatigue, C4 QoL was independent of all other factors. Factors influencing C4 FOS, were C4 fatigue, BL FOS and QoL, with C4 fatigue and BL FOS

exhibiting high consistency (arc strength ≥ 0.93, arc direction ≥ 0.97). Regression estimates (Table 2.3) indicated that worse C4 mood and/or fatigue (i.e., higher score) were associated with worse C4 QoL and FOS.

Factors influencing Y1 QoL and FOS were fewer than at C4. The only predictor for Y1 QoL was Y1 mood, with high consistency (arc strength = 0.96, arc direction = 0.78). C4 FOS and Y1 QoL were directly linked to Y1 FOS but were not stable (arc strength ≤ 0.55).

Several symptoms showed direct cross-sectional and temporal links. At BL and C4, mood was the (only) parent of fatigue with strong cross-sectional links (arc strength ≥ 0.87); also, regression estimates (Table 2.3) were positive, indicating that higher CESD scores (i.e., worse mood) were associated with higher fatigue scores (i.e., worse fatigue). Further, although not exhibiting high consistency, C4 fatigue and sleep quality were parents of Y1 fatigue (arc-strengths ≤ 0.67); Y1 sleep was directly influenced by Y1 mood, and BL sleep (arc-strengths ≤ 0.76). Thus, these links suggest a temporal cluster of sleep, mood and fatigue.

Comparing Networks

Given our focus on cognitive symptoms and quality of life during and after chemotherapy, we conducted sensitivity analyses to test the value of the learned subnetworks for Cognition and QoL. We created a new network in which all edges to- and from- Cognition were removed, and refitted this network to the data. The BIC score for this new network was more than 164 points higher than that of the original networks, indicating far superior fit of the original fitted network and providing support for the identified links to cognition. Similarly, a network in which QoL was isolated (i.e. all edges to- and from- QoL were deleted) had a 226 point higher BIC score, again strongly favoring the original fitted models.

We also tested the impact of removing a specific edge from the network as follows. If we removed the:

- Mood1-QoL1 edge, BIC increased by 12, giving strong evidence for this link

- Fatigue1-QoL1 edge, BIC increased by 2.2, giving weak evidence for this link
- Sleep1-Cognition1 edge, BIC increased by 4.8, giving moderate evidence for this link

Finally, we evaluated the overall fit of our learned network (Figure 2.1). Our a priori assumption was that network structure might vary during chemotherapy versus during follow-up. Hence we fit a flexible network in which links amongst symptoms could be different in the chemotherapy treatment (BL-C4) phase compared to during follow-up (C4-Y1). We can quantitatively assess this assumption via Bayesian information criteria (BIC) scores. Our fitted BL-C4-Y1 network had a BIC score of 4384.4. We then fit a second network in which we constrained the C4-Y1 subnetwork to be identical to the corresponding chemotherapy treatment phase sub-network (the learned BL-C4 sub-network). This constrained network had a BIC score of 4461.6, a 77-point higher score, indicating substantially worse fit for the constrained model compared to the original network.

2.5 Discussion

Most patients undergoing chemotherapy complain of symptoms such as fatigue, impaired sleep and poor mood. Studies of these patients generally focus on average effects and note that mean scores for each of these symptoms usually worsen during chemotherapy. Not much is known regarding how these symptoms influence each other. In the current work, we aimed to address this gap. We applied a powerful Bayesian network approach to discern inter-relationships amongst these symptoms and furthermore, examined the role of these symptoms on QoL, functional outcomes of sleepiness, and cognitive functioning. Unraveling inter-relationships amongst these many factors is a complex computational problem, and Bayesian networks provide a first glimpse at how we might decompose this large multivariate distribution into a set of lower-dimensional relationships.

Clinical Implications

There are many potential clinical implications of this work. Understanding cross-sectional and longitudinal inter-relationships amongst symptoms, QoL and cognition could guide the design of effective interventions. For instance, our networks identified sleep quality as the primary symptom influencing cognition. Thus an intervention aimed at improving sleep during chemotherapy, could potentially mitigate some of the neurocognitive symptoms experienced by cancer patients. We emphasize that our goal in the current analysis was not to assess whether cancer patients experienced cognitive dysfunction, a phenomenon that has been well-studied, but rather to identify factors that might influence acute cognitive ability for a patient undergoing chemotherapy. Another finding of this work was that mood and fatigue directly influenced QoL and/or FOS in the chemotherapy period, and after accounting for the symptom cluster of sleep, mood and fatigue, cognition was (conditionally) independent of QoL and FOS. Thus, an intervention aimed at improving this symptom cluster and implemented while patients are undergoing chemotherapy could have numerous benefits [HRP+15].

Using Bayes information criteria, we were able to confirm our hypothesis that inter-relationships between symptoms and outcomes would be different in the chemotherapy treatment phase (BL-C4) as compared to post-chemotherapy (C4-Y1). It is interesting to note similarities and differences between these networks. Post-chemotherapy quality of life and year 1 quality of life were each influenced by sleep at the same time-point, but post-chemotherapy quality of life was also strongly influenced by concurrent mood, which was not the case for year 1 quality of life, suggesting that interventions to improve mood during chemotherapy could improve post-chemotherapy quality of life. Similarly, while prior functional outcomes of sleepiness (FOS) score influenced subsequent level at all time-points, post-chemotherapy FOS score was strongly influenced by concurrent fatigue, again suggesting that an intervention to reduce fatigue during chemotherapy could improve functional outcomes of sleepiness in breast cancer survivors. Additional differences between the networks are evident in Figures 2.1, but these differences were

not reliable (arc strengths of these differing edges were < 0.7).

Strengths and Limitations

The strengths of this work include a well-characterized cohort of patients undergoing chemotherapy, the use of bootstrap methods and model averaging, which reduce overfit and improve replicability. There are also limitations. Our sample-size is modest and may have impeded our ability to discern important links. Also, our study did not collect self-reported pain, a factor that could have influence the observed findings. Furthermore, Bayesian networks are inherently exploratory. Hence these results need to be confirmed in other cohorts with larger sample-sizes that include broad symptom inventories, including pain, and implement alternative computational strategies such as moderated regression [Fra15a] [Fra15b]. We used an established NP battery which affords the opportunity to evaluate objective cognitive performance during chemotherapy. However, self-reported cognitive deficits are commonly noted by cancer patients during treatment. It would be interesting to investigate if networks for self-reported versus objective cognition are similar, and we leave this question and other similar ones (e.g., comparing objective sleep assessed via actigraphy to self-reported sleep) for a future study.

2.6 Conclusions

In this article, we have introduced Bayesian networks, a machine learning methodology, to infer networks of symptom cluster and cognitive and psychosocial outcomes for breast cancer patients during and one-year after undergoing chemotherapy. Our results identified separate pathways and potential links between symptoms, cognitive function and QoL. The network comparison analysis strongly favored the fitted networks, indicating that our findings are robust against alternative network structures. Our work illustrates that Bayesian networks could be a powerful tool in cancer symptoms research; we advocate their use in future studies.

2.7 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in “Cognition, qualityoflife, and symptom clusters in breast cancer: Using Bayesian networks to elucidate complex relationships”, *PsychoOncology*, 2017. Selene Yue Xu, Wesley Thompson, Sonia Ancoli-Israel, Lianqi Liu, Barton Palmer, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Table 2.1: Participant characteristics (N=74, female breast cancer survivors)

	Mean (SD) or percent
Age (years)	51 (9.5)
Education	
High-school	11%
Some college	38%
College graduate	51%
Race	
Caucasian	88%
Asian	7%
African-American	3%
Ethnicity	
Hispanic	8%
Cancer Stage	
Stage I	27%
Stage II	41%
Stage III	31%

Table 2.2: Symptom scores (mean (SD)) for 74 breast cancer patients before, at completion of, and 1 year after chemotherapy treatment

Symptom	Direction of better outcome	Pre-chemotherapy	End-of chemotherapy-cycle4	One-year post-chemotherapy
Cognition (NP composite score)	↑	0.062 (0.743)	0.077 (0.691)	0.166 (0.738)
Mood (CESD)	↓	11.5 (10.4)	16.2 (12.9)	10.0 (9.95)
Fatigue (MFSI)	↓	9.66 (18.3)	18.0 (23.9)	7.6 (20.3)
Quality of Life (FACT-B)	↑	105 (16.1)	95.0 (23.3)	110 (19.0)
Sleep quality (PSQI)	↓	7.71 (3.87)	9.07 (3.74)	7.49 (4.40)
Functional outcomes of sleepiness (FOS)	↑	18.0 (2.04)	16.0 (2.87)	17.7 (2.19)

Table 2.3: Bayesian network structure and associations. Suffixes: 0=pre-chemotherapy; 1= end-of-chemotherapy (1); 2= one year after the start of chemotherapy

Outcome (child)	Predictors (parents)	strength	direction	regression coefficients (SE)
Mood1	QoL0	0.52	0.94	-0.536 (0.095)
Cognition0	Age	0.86	1.00	-0.043 (0.010)
Cognition1	Cognition0	1.00	1.00	0.857 (0.062)
	Sleep1	0.68	1.00	-0.037 (0.012)
Cognition2	Cognition0	0.95	1.00	0.510 (0.113)
	Cognition1	0.93	0.90	0.494 (0.121)
QoL0	Fatigue0	0.93	0.72	-0.706 (0.084)
QoL1	Mood1	0.94	0.89	-0.980 (0.237)
	Fatigue1	0.64	0.58	-0.290 (0.121)
	QoL0	0.63	0.99	0.242 (0.110)
QoL2	Mood2	0.96	0.78	-1.583 (0.171)
FOSQ0	Fatigue0	0.70	0.64	-0.069 (0.014)
FOSQ1	Fatigue1	0.95	0.97	-0.097 (0.012)
	QoL0	0.62	1.00	-0.047 (0.019)
	FOSQ0	0.93	1.00	0.582 (0.131)
FOSQ2	QoL2	0.55	0.64	0.057 (0.013)
	FOSQ1	0.50	0.99	0.329 (0.085)
Fatigue0	Mood0	0.99	0.69	1.400 (0.169)
Fatigue1	Mood1	0.87	0.90	1.651 (0.135)
Fatigue2	Fatigue1	0.62	0.99	0.169 (0.112)
	QoL2	0.73	0.64	-0.520 (0.132)
	Sleep1	0.67	0.97	1.439 (0.677)
Sleep0	QoL0	0.96	0.73	-0.158 (0.029)
Sleep1	Fatigue1	0.67	0.85	0.080 (0.020)
	Sleep0	0.52	0.99	0.285 (0.125)
Sleep2	Mood2	0.75	0.87	0.219 (0.050)
	Sleep0	0.76	1.00	0.353 (0.116)
	FOSQ2	0.53	0.57	-0.457 (0.227)

Table 2.4: Summary statistics (Mean (SD)) for Neuropsychological test battery

NP Test	Pre-chemotherapy	End-of-cycle 4	Year 1
	Mean (SD)	Mean (SD)	Mean (SD)
HVLT TotalRecall	29.7 (4.12)	30.46 (3.81)	31.54 (3.83)
WAIS DigitSpan	18.45 (3.69)	18.64 (4.07)	18.66 (3.95)
WAIS DigitSymbol	77.51 (18.87)	80.00 (19.51)	83.15 (18.75)
WAIS SymbolSearch	34.54 (8.42)	36.7 (8.73)	39.55 (9.93)
TRAILS A	29.03 (15.41)	26.13 (10.55)	23.9 (10.68)
TRAILS B	60.67 (45.61)	52.98 (27.48)	54.17 (41.05)
WCST Concept	42.55 (12.57)	47.49 (10.5)	47.93 (10.9)
STROOP Correct Word	104.12 (17.88)	102.14 (16.44)	100.66 (22.46)
STROOP Correct Color	77.04 (12.37)	75.05 (14.12)	77.29 (13.98)
STROOP CorrectColorWord	44.99 (9.8)	45.78 (10.84)	46.61 (9.98)
LACF Total	48.00 (13.47)	47.97 (12.77)	50.62 (15.15)
LACF Animal	22.94 (5.89)	22.88 (6.05)	22.95 (5.84)
Digit Cancel	32.2 (5.38)	31.36 (6.27)	32.85 (5.11)

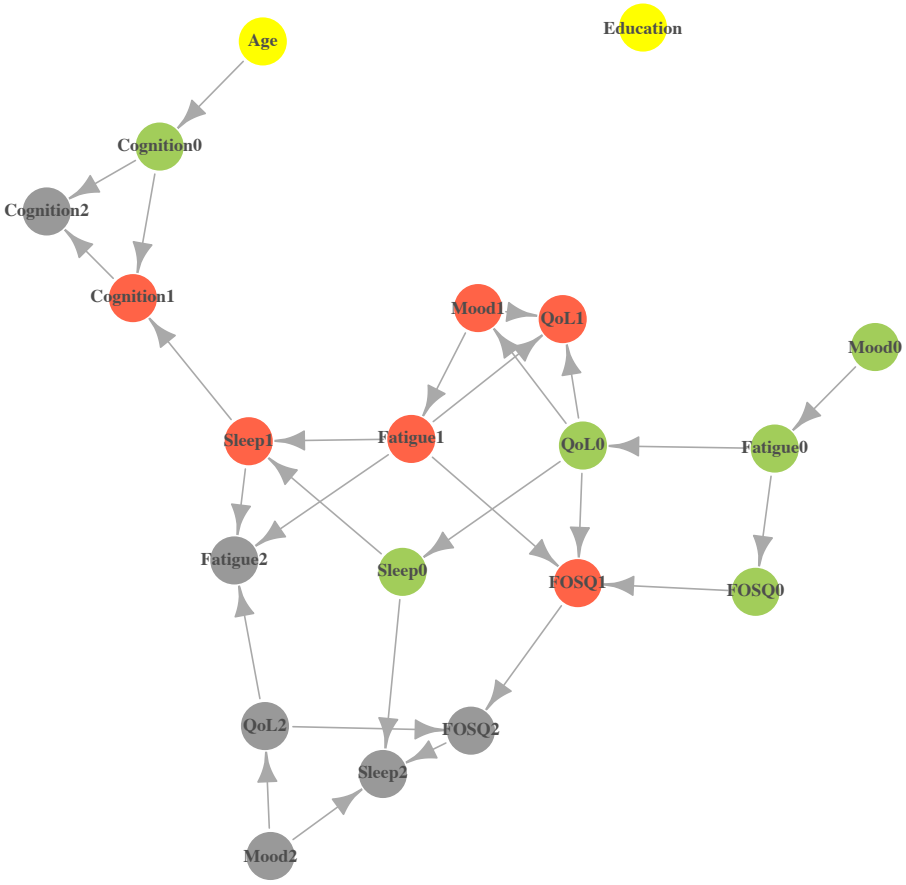

Figure 2.1: Bayesian network of symptoms and outcomes before (BL), during (C4) and after chemotherapy (Y1) among breast cancer patients. The circles represent variables and arrows codify dependencies between variables. The colors and number suffixes represent the three time periods: green, 0=BL; red, 1=C4; grey, 2=Y1

Chapter 3

Bayesian Network Project - Reach for Health

3.1 Introduction

Obesity, physical inactivity, and impaired sleep are known risk factors for cardiovascular disease (CVD) and cancer [BBGAS13] [XKHM14] [PDRB14] [CIMS15] [Par14] [LAC+14] [VB02] [CRWTT03]. Sedentary behavior is also reported to increase risk for disease and mortality [LS16] [BOF+15]. It is increasingly recognized that multiple biological pathways, such as glucose regulation and inflammation, are implicated in the link between health behaviors, obesity and chronic disease [NP15] [CWC15]. Unraveling interrelationships among these factors could elucidate disease mechanisms, and inform design of clinical studies.

Modeling multiple correlated factors can be computationally challenging and requires new statistical approaches. Standard regression modeling cannot disentangle these complex associations. Bayesian graphical networks are a novel and powerful approach for examining relationships among multiple correlated variables. These models provide algorithms for discovering and analyzing structure, as well as, an intuitive graphical interface for visualizing multivariate

distributions. Bayesian networks were initially developed in computer science and artificial intelligence applications [Pea14] [Pea09], and have since had major impact in biomedicine, in omics studies, neurosciences, and more recently in obesity research [ZSAK$^+$06] [YRL14] [KWB$^+$13] [FPTBV$^+$16] [FLNP00] [FBB$^+$11] [ESV$^+$15] [BL14] [ATdMR10]. To our knowledge, no existing obesity studies have examined multiple lifestyle factors simultaneously in conjunction with biomarkers and psychosocial factors in a cancer population.

In this work, we applied probabilistic Bayesian networks [Pea09] [NSL13] to elicit bio-behavioral pathways implicated in obesity and health in breast cancer survivorship. Our sample comprised 333 well-characterized postmenopausal breast cancer survivors with objective physical activity assessments, and detailed information on multiple lifestyle factors, clinical characteristics, and health measures. Using this unique sample, we developed Bayesian networks to model inter-relationships between health behaviors (sleep, physical activity), body mass index (BMI), biomarkers of inflammation (C-reactive protein [CRP]) and glucose regulation (insulin), and mental and physical quality of life.

3.2 Brief description of Bayes network methodology

A Bayesian network is a probabilistic directed acyclic graph. Random variables are depicted as nodes on the graph, and edges between nodes represent dependencies (e.g., partial correlations) between these variables. If there is a directed link (arrow) from node A to node B, then A is termed the "parent" and B the "child". Each node has an associated distribution function that takes as input a set of values for the node's parent variables and gives the probability of the variable represented by the node. The presence of an edge or path between two variables indicates a non-zero partial correlation between the two variables.

Fitting a Bayesian network requires (i) learning its structure, namely which nodes in the graph are connected, and (ii) estimating parameters associated with conditional probabilities.

Specifically, let X comprise the set of variables X_i (e.g., X_1 = physical activity, X_2 = sleep quality, X_3 = BMI, X_4 = insulin level, etc.) and M be a Bayesian network on X, comprising a directed acyclic graph of edges between variables in X. The model M encodes conditional independencies that imply a factoring of the joint probability distribution $p(X)$ of X [Pea09]:

$$Pr(X|M) = \prod Pr(X_i|pa(X_i)) \tag{3.1}$$

where $pa(X_i)$ denotes variables ("parents") in X with arrows leading into X_i. The structure of the graph M can be learned by implementing constraint-based, search-score, and hybrid algorithms [SGS00]. For a given graph M, $Pr(X_i|pa(X_i))$ represents a local probability distribution, and its parameters β can be estimated by regression methods, using multivariate Gaussian distributions (after appropriate transformation if needed) or non-parametric approaches for continuous variables and multinomial distributions for categorical variables [NSL13] [N$^+$04]. Thus, dependencies and (conditional) independencies between sets of variables can be derived from a Bayesian network analysis.

The notion of a Markov blanket of a Bayesian network can be used to identify sets of variables that are (conditionally) independent (i.e., uncorrelated). The Markov blanket for a node V in a Bayesian network is the set of nodes composed of Vs parents, children and its childrens other parents. The node V is conditionally independent of any set of nodes in the network when conditioned on its Markov blanket [NSL13].

3.3 Materials and Methods

3.3.1 Ethics Statement

This was a secondary data analysis of the "Reach for Health" clinical trial carried out at the University of California, San Diego (UCSD). The original study was approved by the UCSD

IRB board, project # 101977. All subjects in the Reach for Health study provided written consent.

3.3.2 Study Sample and Measures

Study sample

Our study sample comprised 333 early-stage breast cancer survivors enrolled in a weight-loss intervention. Details regarding the study protocol and design have been previously published [PMN+15]. Briefly, the study enrolled breast cancer survivors, who were postmenopausal at cancer diagnosis, were either overweight or obese at study entry, and had completed primary breast cancer treatment (surgery with or without chemotherapy and radiation). 83% were white; 11% were Hispanic. More information on demographics, lifestyle, clinical factors, coping, sleep, mood, physical factors, and biomarkers is provided in Table 3.1. The current analysis used baseline information to develop network models.

Measures

We obtained participants medical records including tumor characteristics (Cancer Stage, hormone receptor status) and years from cancer diagnosis to study entry (YrsDXRND). During clinic visits, participants height and weight were measured and used to calculate BMI. Physical activity (PA) and sedentary behavior (SB) were determined by objective 7-day, minute-level triaxial accelerometer counts. Specifically, PA was the average (across days) of total counts per minute per day, thus representing a measure that captured total volume of activity; moderate vigorous physical activity (MVPA) was the average of minutes per day with counts ≥ 1952; SB was the average of minutes per day with counts < 100. Accelerometer-derived measures were adjusted for device wear-time. Demographic information and other study measures were obtained through self-report or questionnaires. The Neighborhood Environment Index (Neighborhood) derived from the NEWS scale [CSSF06] was used to measure walkability. It has a range from 0

to 6, with higher scores indicating more walkable neighborhood. Sleep quality was evaluated based on the PROMIS scale [BYM+10]. In the current analysis, we used two subscales, the sleep disturbance (sleep1), and the sleep impairment (sleep2) subscales. These subscales were normed to mean 50 with standard deviation of 10. Higher scores indicated worse sleep. Quality of life assessment, both mental (QOLm) and physical (QOLp), used the SF-36 scale [BHJ+92]. QOLm and QOlp scores from 0 to 100, with higher scores reflecting better quality of life. The Monitor-Blunter (MB) scale assessed participants coping mechanism. It ranged from -16 to 16, with higher scores indicating more monitor than blunter. Fasting plasma CRP and insulin concentrations were measured using immune-based assays (Meso Scale Discovery).

3.3.3 Statistical Methods

We fit a Bayesian network to examine multivariate relationships between demographics, clinical factors, health behaviors and health outcomes. We disallowed implausible edge directions while learning the network structure. Specifically, we disallowed QOLp and QOLm to be the parent nodes of any other variable in the network; and we disallowed age, education, cancer stage, years between diagnosis and study entry, and neighborhood to be the child nodes of any other variable. We applied bootstrap resampling to learn a set of 500 network structures. We then averaged these networks in an attempt to reduce the impact of locally optimal (but globally suboptimal) networks on learning and inference. The averaged network is a more robust model with better predictive performance than choosing a single, high-scoring network [NSL13]. To quantify stability of inferred edges, we computed arc strength and direction strength. Arc strength was calculated as the frequency of an edge occurring between two variables across the 500 bootstrapped network structures; similarly, directional strength was assessed as the frequency of the observed direction re-occurring in the set of learned network structures in which the relevant edge occurred. The averaged network was created using the arcs whose strength exceeded a threshold, which was computed by searching for the arc set "closest" to the arc strength computed

from the original data [NSL13]. Conditional independencies were inferred using Markov blankets and related Bayesian network theory.

We used Bayesian information criteria (BIC) and posterior model probabilities to compare fit of candidate networks. The BIC was computed as $logLik(M) - 0.5 \times k \times log(n)$, where $logLik(M)$ is the log-likelihood of model M, k is the number of parameters in M, n is the sample-size. This is the classic definition rescaled by -2; hence, in our calculations, higher BIC scores indicate better fit. We also calculated the Bayes factor, which is the ratio of the posterior probabilities (given the observed data) of the first to the second model, as another metric to compare the two models. The log of the Bayes factor can be approximated as the difference in the BIC scores as defined above [Was00]. Biomarkers were log-transformed to better approximate Gaussian assumptions. Models were fitted using the R package bnlearn [NSL13].

3.4 Results

3.4.1 Decomposition of probability distribution

The fitted network is shown in Fig 3.1. From the network analysis, we can obtain the joint probability distribution of all the variables as a product of conditional distributions. In our application, we obtained:

$$P(X) = P[CancerStage] \times P[YearsDXRND] \times P[Neighborhood] \times P[Drink]$$
$$\times P[MB] \times P[Education] \times P[Age] \times P[Smoke] \times P[Insomnia]$$
$$\times P[Depression|Insomnia] \times P[Sleep1|Insomnia + Depression] \times P[Arthritis|Depression]$$
$$\times P[BMI|Smoke + Arthritis] \times P[Sleep2|Depression + Sleep1]$$
$$\times P[QOLp|BMI + Sleep2 + Arthritis] \times P[QOLm|Depression + Sleep2]$$
$$\times P[Insulin|BMI] \times P[CRP|BMI] \times P[PA|Age + Sleep2 + Insulin] \quad (3.2)$$

This decomposition converts the complex model comprising 19 variables into simpler components, and highlights subsets of factors that directly influence each variable. In fact, the maximum number of directed edges pointing to any variable is 3 (e.g., PA and QOLp), substantially fewer than the maximum of 18 possible directed edges. A first notable finding is that there were no edges from (or to) the following variables: cancer characteristics (stage, years from diagnosis to study entry), neighborhood, education, alcohol intake and coping style (MB scale), indicating that these variables were (marginally) independent of all other factors. Below, we provide additional details on these decompositions, and how to infer (in)dependencies between variables.

3.4.2 BMI and physical activity (PA)

The network allows us to elicit local structure, so that we can identify "parents", namely variables that directly influence any given factor. In our learned network (Fig 3.1), the variables smoke and arthritis were parents of BMI. Table 3.2 provides parameter estimates, and strength of network links based on bootstrap analysis. The smoke and arthritis links to BMI were not very stable as reflected in the low arc-strengths from the bootstrap analysis: 0.63 for the smoke-BMI

link, 0.54 for the arthritis-BMI link (Table 3.2). The regression coefficient for arthritis was positive, indicating that having arthritis was associated with higher BMI on average. Interestingly, the regression coefficient for smoke was positive as well, indicating that smoking was associated with higher BMI, which is contrary to the common belief that smoking can cause weight loss by suppressing appetite. However, smoking status in our cohort refers to "ever smoking" and likely reflects former smokers who quit many years ago; the proportion of current smokers was $<2\%$, too few to include as a separate variable in our analysis. In addition, BMI had a large Markov blanket comprising smoke, sleep2, QOLp, arthritis, insulin, and CRP, indicating its influence on multiple factors. Also, using the theory of Bayesian networks, we can infer that BMI is independent of all other variables conditional on its Markov blanket.

Next, we examined links to PA. Age, sleep impairment (sleep2), and insulin level were parents of PA (Fig 3.1). The association between age and PA had the highest arc strength (0.97), with a negative regression coefficient showing that higher age was associated with, on average, lower level of physical activity. The link between insulin and PA had a moderately-high arc strength (0.73) and directional strength (0.77), whereas the link between sleep2 and PA was weak (arc strength = 0.53) but had a relatively strong directional strength (0.80). The regression coefficients for insulin and sleep2 were negative, implying that higher insulin level and poor sleep were associated with lower physical activity level. Since PA did not have any children in our learned network, the parents of PA also comprised the Markov blanket of PA. Thus, once we observe a subjects age, sleep2, and insulin, her physical activity level is independent of all other variables in the network. It is interesting to note that BMI had a strong positive association with insulin (arc strength = 1.00, directional strength = 0.82). Hence, we can infer that BMI was indirectly negatively associated with PA.

3.4.3 Biomarkers (insulin and CRP)

We are also interested in studying the local network structure of the two biomarkers, insulin and CRP. Both markers shared a single parent, BMI, and for both, this link appeared in 100% of the bootstrapped networks (arc strength = 1.00). The link between BMI and CRP also had very strong directional strength of 0.96, with a moderately high value of 0.82 for the BMI-insulin link. Both regression coefficients were positive, so that higher BMI was associated with higher insulin and CRP. The Markov blanket for insulin consisted of BMI, age, sleep2, and PA; and the Markov blanket for CRP only had only one element, BMI.

3.4.4 Quality of life (physical and mental)

We also briefly summarize interesting associations revolving around physical and mental quality of life (QOLp & QOLm). BMI, sleep2 and arthritis were parents of QOLp (Fig 3.1). Both BMI and arthritis had strong associations with QOLp, with arc strength of 0.83 and 0.80 respectively. Regression coefficients showed that QOLp was, as expected, negatively associated with both BMI and arthritis (Table 3.2). We note that arthritis was directly, and indirectly via BMI, linked to QOLp, implying that BMI could be a mediator between arthritis and QOLp. Surprisingly, sleep2 had the strongest association with QOLp (arc strength = 1.00), with a corresponding negative regression coefficient indicating that poor sleep quality was associated with worse physical quality of life.

Depression and sleep2 were parents of QOLm, with respective arc strengths of 0.95 and 1, indicating that this cluster was strongly linked and highly reproducible. Again, as expected, the negative regression coefficients suggested that poor sleep and depression were associated with poorer mental QoL. Finally, via Markov blankets we infer that, conditional on BMI, sleep2, and arthritis, QOLp was independent of all other factors; and, conditional on depression and sleep2, QOLm was independent of all other variables.

3.4.5 Hubs and subnetworks

In our network (Fig 3.1), the set (or any subset) of variables {insomnia, depression, QOLm} was conditionally independent of the set (or any subset) of variables {BMI, PA, QOLp, insulin, CRP}, given {sleep2, arthritis}. We point out that arthritis was in the set of conditional variables due to its link to depression, however, the arthritis-depression link was in fact weak with arc strength of 0.69 and even weaker directional strength of 0.64. This implies that sleep quality was the primary hub linking mental factors to physical health and biomarkers.

3.4.6 Comparing Networks

Given the finding that sleep played a central role in our networks, we conducted network comparison analyses to test the importance of the two sleep quality measurements, sleep1 and sleep2. We quantitatively assessed this assumption via Bayesian information criteria (BIC) scores. The original learned network had a BIC score of -14483.5 We then fit a second network by isolating sleep1, i.e., removing all links to and from sleep1, and obtained a BIC score of -14637.4, a 154-point lower score, indicating substantially worse fit for the model with the sleep1 variable isolated compared to the original network. The Bayes factor for the original vs second model was approximately exp(-14483.5+14637.4), indicating $>$ 20-fold higher posterior probability for the original compared to the sleep-omitted network, thus affirming our hypothesis that sleep1 plays a critical role in the network. Similarly, isolating sleep2 resulted in an even larger reduction of 190 points in the BIC score, and hence a Bayes factor that strongly favored the original model. These analyses confirm the role of sleep as an important factor in the fitted network.

Given our focus on BMI and biomarkers, we conducted additional network comparison analyses to test the value of the learned sub-networks for BMI and the two biomarkers. We created a new network in which the edge from BMI to insulin was removed. The BIC score for this network was 20 points lower than that of the original network, and, as before, the Bayes

factor would strongly favor the original model. Similarly, a network in which the edge from BMI to CRP was removed had a 26 points lower BIC score, again strongly favoring the original fitted model.

Finally, we investigated the depression-arthritis link, which was reproduced in only 69% of bootstrapped networks. Omitting this link, decreased the BIC score by 1.56, indicating moderate evidence for this association.

3.4.7 Deconstructing total PA

To further investigate physical activity, we parsed the total PA volume (counts/minute) variable as two activity behaviors: sedentary time and MVPA. When these two "activity" variables were included in the network instead of total PA, the network structure and parameters were almost identical to the original network (Table 3.3, Fig 3.2). The only change in structure for the non-PA variables was that the edge between arthritis and BMI was dropped, and age was isolated and independent of all variables. With regards to activity, MVPA and arthritis were both direct parents, as well as, the Markov blanket of sedentary time, with lower MVPA and having arthritis associated with more sedentary time. The MVPA-sedentary time link was reproduced in 100% of the bootstrapped networks, and the network in which this link was omitted had a 28 point lower BIC score. The arthritis-sedentary time link was less robust occurring in 63% of bootstrapped networks with a corresponding 2.16 lower BIC score when this link was dropped.

3.4.8 Predicting Intervention Effects

Edges and paths inferred from a Bayesian network can be used for prediction. For instance, if we perturb a node, say PA or BMI, we can investigate predicted downstream effects on biomarkers. Table 3.4 gives a few examples of such queries: increasing total PA from $<$270 count/min/day to $\geq$ 380 counts/min/day (i.e., a PA increase of 1 SD) would be predicted to result

in an average BMI reduction of 2.3 kg/m^2. Or, moving from the obese to overweight category would be predicted to reduce insulin by 27% (reduction in loginsulin is 0.32pg/ml), reduce CRP by over 50% (reduction of 0.75 mg/l in logCRP), improve physical QoL by an average 6 points, but to not change mental QoL appreciably. Most interestingly, reducing sleep impairment from the highest to the lowest quartile would be predicted to improve both mental and physical QoL by > 20 points. And, a combined change of reducing BMI category from obese to overweight and reducing sleep impairment from highest to lowest quartile, would be predicted to result in a 26-point higher physical QoL score on average, suggesting that an intervention aimed at weight-loss and reducing sleep impairment could have additive effects on physical QoL.

3.5 Discussion

In this work, we have illustrated how Bayesian networks, a machine learning tool, can be applied in behavioral research. Health behaviors are modifiable risk factors, and hence can be potentially intervened upon to improve health and reduce disease. Nevertheless, identifying which behaviors are most robustly linked to disease is critical for designing effective interventions. Bayesian networks can shed light on this question, as we enumerate below.

1. Identifying intervention targets: Bayesian networks provide insights into which factors directly affect health. For instance, in our analysis, BMI was directly linked to the biomarkers, suggesting that a weight-loss intervention could improve profiles of these markers. Similarly, sleep impairment was directly linked to Quality of life (mental and physical) suggesting that an intervention aimed at improving sleep quality could improve QoL. Of note, our network also suggests that a combined sleep improvement and weight-loss intervention could improve physical and mental QoL, as well as, glucose regulation and inflammation.

2. Mechanisms: Bayesian networks can identify indirect pathways of influence. For example, the arthritis-BMI-QoLp link indicates that high BMI is one of the mechanisms by which

arthritis impacts QoL. Similarly, the depression-sleep2-QoLm link identifies sleep impairment as an intermediate factor by which depression impacts mental health. Again, these indirect paths through health behaviors suggest intervention targets, namely weight and sleep2, that could reduce the impact of arthritis and depression on physical and mental QoL respectively.

3. Informing study design: As shown in Table 4, Bayesian networks can be used to estimate putative intervention effects, and hence inform achievable effect-sizes and required sample-size.

4. Tailoring interventions: Bayesian networks can be useful for identifying at-risk populations and personalizing interventions. For instance, our network indicates that older age, more sleep impairment and higher BMI are each associated with lower physical activity, suggesting that these three factors could be used to streamline a proposed physical activity intervention to be most responsive to the needs of specific subgroups.

We have enumerated a few ways in which Bayesian network analyses could inform public health research. The strengths of this work include a well-characterized cohort of breast cancer survivors, the availability of clinical information from medical records, objective information on physical activity, biomarker outcomes, and from a methodological perspective, the use of bootstrap methods and Bayesian information criteria, which reduce overfit and improve replicability. There are also limitations. Bayesian networks are an inherently exploratory tool, best suited for hypothesis generation. Hence our results need to be confirmed in other cohorts and/or randomized trials. Also, our cohort only included overweight postmenopausal cancer survivors who agreed to participate in an intervention trial, which could limit generalizability. For instance, it is possible that with an unrestricted BMI range, we may have observed other factors (e.g., built environment, age, PA) influencing BMI and other outcomes. Nevertheless, it would be interesting to test our final averaged network on younger and/or normal weight breast cancer survivors.

3.6 Conclusion

In conclusion, in this article, we have introduced Bayesian networks, a machine learning methodology, to infer behavioral networks in a breast cancer cohort. Our results identified several health behaviors directly linked to biomarker and quality of life outcomes, suggesting potential mechanistic pathways and useful intervention targets. The network comparison analysis strongly favored the fitted networks, indicating that our findings are robust against alternative network structures. We believe that this network methodology could be a useful tool in health behaviors research.

3.7 Acknowledgements

Chapter 3, in full, is a reprint of the material as it is under review in "Modeling interrelationships between health behaviors: applying Bayesian networks". Selene Yue Xu, Wesley Thompson, Jacqueline Kerr, Suneeta Godbole, Dorothy Sears, Ruth Patterson, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

Table 3.1: Characteristics of the Study Cohort (N = 333)

	Variables (Nodes)	Mean (SD)/Percentage
Demographics & Lifestyle	Age	63.1 (6.9) yrs
	Education	51% had college degree or higher
	Smoke (Ever)	45% Yes
	Alcohol	median 4 drinks/month
Clinical Factors	Cancer Stage	48% Stage1; 35% Stage2; 17% Stage3
	Time: cancer diagnosis to randomization	2.7 (2.0) yrs
	Estrogen Receptor	85.0% positive
	Progesterone Receptor	71.8% positive
Cancer Treatment	Treatment type	53.2% chemotherapy; 72.1% radiation; 76.9% endocrine; 13.8% immunotherapy
Coping	Monitor-blunter (MB)	4.2 (3.5)
Neighborhood	NEWS scale	3.1 (1.7)
Health	Insomnia	28.8% Yes
	Depression	40.8% Yes
	Arthritis	56.4% Yes
	QOLp (SF-36)	66.2 (18.7)
	QOLm (SF-36)	73.6 (18.4)
	Insulin	median (25th, 75th)-%ile 463.8 (335.6, 667.0) pg/ml
	CRP	median (25th, 75th)-%ile 3.1 (1.5, 6.5) mg/l
Health Behaviors	Physical Activity (PA)	273.1(108.6) counts/min/day
	MVPA	17.5 (17.3) min/day
	Sedentary time	471(111) min/day
	BMI	31.1(4.9) kg/m^2
	Sleep Disturbance (Sleep1)	50.5 (8.8)
	Sleep Impairment (Sleep2)	46.9 (9.0)

Table 3.2: Parameter estimates and stability: network of total physical activity, sleep, BMI, biomarkers, and psychosocial functioning. *: coefficients represent log-odds ratios. +: sleep1=sleep disturbance, sleep2=sleep impairment

Outcome (child)	Predictors (parents)	Strength	Direction	Regression coefficients (SE)
Arthritis*	Depression	0.69	0.64	0.704 (0.239)
BMI	Smoke	0.63	1.00	1.157 (0.551)
	Arthritis	0.54	0.98	1.524 (0.555)
CRP	BMI	1.00	0.96	0.093 (0.012)
Depression*	Insomnia	0.93	0.60	1.000 (0.257)
Insulin	BMI	1.00	0.82	0.039 (0.006)
PA	Age	0.97	1.00	-4.563 (0.845)
	Sleep2+	0.53	0.80	-1.905 (0.659)
	Insulin	0.73	0.77	-43.926 (11.188)
QOLm	Depression	0.95	0.97	-7.029 (1.717)
	Sleep2+	1.00	0.95	-1.060 (0.094)
QOLp	BMI	0.83	1.00	-0.689 (0.176)
	Sleep2+	1.00	1.00	-0.921 (0.096)
	Arthritis	0.80	1.00	-7.439 (1.740)
Sleep1+	Insomnia	1.00	0.97	10.641 (0.915)
	Depression	0.53	0.83	1.710 (0.845)

Table 3.3: Parameter estimates and stability: network of sedentary behavior, moderate-vigorous physical activity, sleep, BMI, biomarkers, and psychosocial functioning. *: coefficients represent log-odds ratios. +: sleep1=sleep disturbance, sleep2=sleep impairment

Outcome (child)	Predictors (parents)	Strength	Direction	Regression coefficients (SE)
Arthritis*	Depression	0.69	0.62	0.704 (0.239)
BMI	Smoke	0.66	1.00	1.196 (0.557)
CRP	BMI	1.00	0.95	0.093 (0.012)
Depression*	Insomnia	0.94	0.65	1.000 (0.257)
Insulin	BMI	1.00	0.85	0.039 (0.006)
QOLm	Depression	0.93	0.97	-7.029 (1.717)
	Sleep2+	1.00	0.97	-1.060 (0.094)
QOLp	BMI	0.84	1.00	-0.689 (0.176)
	Sleep2+	1.00	1.00	-0.921 (0.096)
	Arthritis	0.79	1.00	-7.439 (1.740)
Sedentary	Arthritis	0.67	0.99	-11.706 (7.178)
	MVPA	1.00	0.90	-1.788 (0.213)
Sleep1+	Insomnia	1.00	0.96	10.641 (0.915)
	Depression	0.53	0.79	1.710 (0.845)
Sleep2+	Depression	0.83	0.96	3.489 (0.741)
	Sleep1+	1.00	0.93	0.665 (0.042)

Table 3.4: Bayesian network prediction. Q3: 75th%-ile; Q1: 25th%-ile

Target	Change in targeted behavior(s)	Outcome	Average change in outcome
PA	$<$ 270 to $\geq$ 380 counts/min/day	BMI	BMI decreases by 2.3 kg/m^2
BMI	$\geq$ 30 to $<$ 30 kg/m^2	Insulin	Insulin decreases by 0.32 (log) pg/ml (27% decrease)
BMI	$\geq$ 30 to $<$ 30 kg/m^2	CRP	CRP decreases by 0.75 (log) mg/l (50% decrease)
BMI	$\geq$ 30 to $<$ 30 kg/m^2	QoLm	QoLm increases 0.5
Sleep2	$>$ Q3 to $<$ Q1	QoLp	QoLp increases 21.6
Sleep2	$>$ Q3 to $<$ Q1	QoLm	QoLm increases 26.9
Sleep2 + BMI	sleep2 $>$ Q3 to $<$ Q1 BMI $\geq$ 30 to $<$ 30 kg/m^2	QoLp	QoLp increases 28.2
Sleep2 + BMI	sleep2 $>$ Q3 to $<$ Q1 BMI $\geq$ 30 to $<$ 30 kg/m^2	QoLm	QoLm increases 26.9

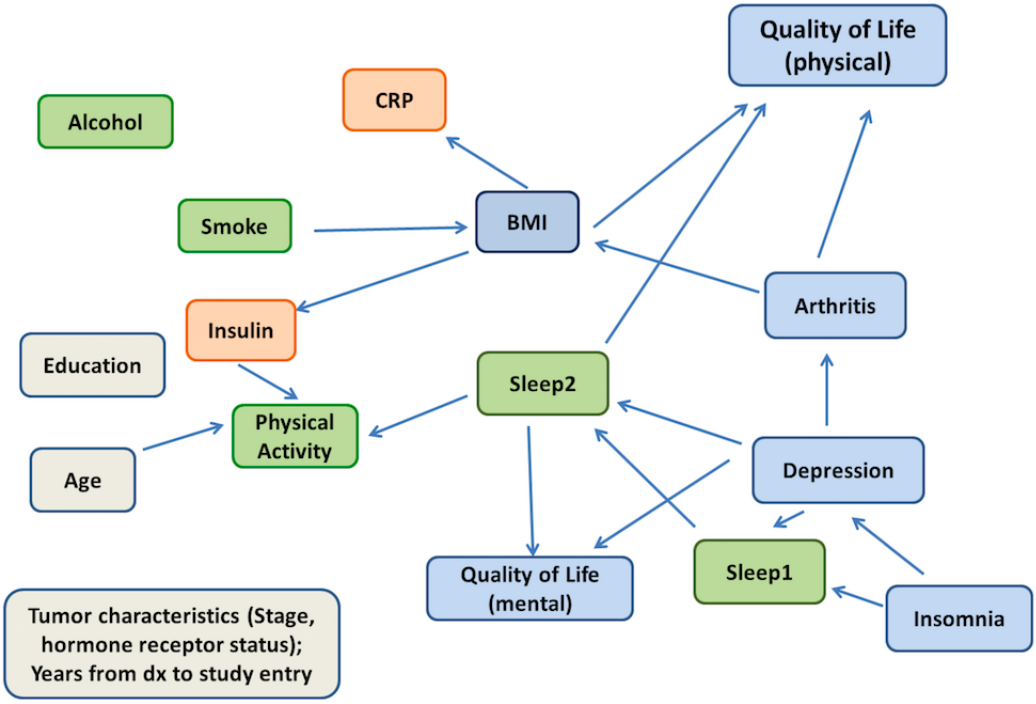

Figure 3.1: Bayesian network of total physical activity, sleep, BMI, biomarkers, and psychosocial functioning

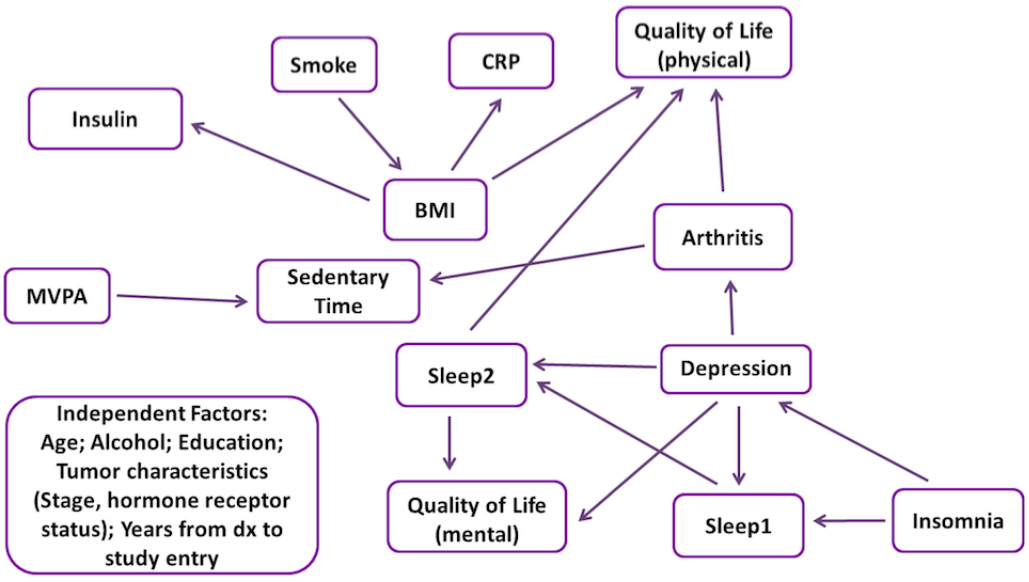

Figure 3.2: Bayesian network of sedentary behavior, moderate-vigorous physical activity, sleep, BMI, biomarkers, and psychosocial functioning

Chapter 4

Functional Method Project - Using Quasi-Poisson Processes to Model Accelerometer Data

4.1 Introduction

Advances in technology have resulted in the use of sensors in a great variety of applications ranging from weather forecasting, GPS tracking to physical activity measurement. Novel analytic techniques are being developed to study these densely sampled data. Our paper introduces a unique framework to model accelerometer data based on sound statistical theory and machine learning techniques.

Accelerometers are wrist- or hip-worn sensors that provide objective measurements of body movement via acceleration counts. More importantly, these devices can continuously monitor movements over time and therefore offer a rich framework for elucidating differing patterns between individuals, as well as, day-day variability within an individual. The majority of research utilizing accelerometers has traditionally focused on aggregated statistics such as

daily average activity or percentage of sedentary/vigorous activity. More recently functional analysis techniques such as wavelet-based or principal component functional models [MC06] [MAC+06a] [XHS+14] have been proposed to study the full spectrum of accelerometer derived activity profiles. These functional techniques have offered important insights into physical activity and public health among various demographic groups ranging from children to the elderly, and even among other species such as the domestic cat [GLJR16] [GACT+17] [SGD+17] [ZBH+17]. Furthermore, efficient algorithms have been developed to accomodate complex sampling designs and longitudinal effects [SZCG15] [LSC14] [LKKS+15] [LKKC16] [LSW+18]. While these data-driven methods hold great promise for parsing accelerometer data, we propose a theory-driven approach based on non-homogeneous Poisson processes and utilizing multi-level functional principal component analysis techniques [DCCP09]. To our knowledge, the classical theory of Poisson processes has never been applied to the field of accelerometry. We believe our model could provide a more refined statistical understanding of the underlying dynamics of human physical activity and a systematic computational approach to analyze accelerometer data incorporating subject-specific information and study design features.

In this article, we leverage minute-level accelerometer data from two behavioral intervention studies conducted at the School of Medicine at the University of California San Diego (UCSD): MENU a randomized diet intervention among overweight women [LFN+16] and Reach for Health a trial examining the use of metformin and a lifestyle intervention in overweight, breast cancer survivors [PMN+16]. These two trials provide us with accelerometer data on 578 overweight women, as well as a rich array of health status measurements including biomarkers and quality of life. We not only build our model on this data set but also showcase how model-derived parameters can provide novel information on physical activity patterns and health.

The next section provides the theoretical foundation of our model. This consists of two main parts: 1. determine the underlying activity trend as the intensity function in a non-homogeneous Poisson process 2. account for over-dispersion by expanding on the base model.

Section 4.3 then illustrates some of the empirical results as we fit the model on our data. This includes statistical analysis of health outcomes in association with the physical activity patterns learned by our model. Section 4.4 introduces two metrics on which we evaluate our model and compare the performance of the different variations of the model proposed in the paper.

4.2 Statistical Framework

Accelerometer counts form irregular functions characterized by peaks of varying frequencies and locations, as illustrated for a randomly chosen subject in Figure 4.1. On this day, this subject wore the accelerometer for 863 minutes from 8:22am to 10:44pm. Here we give a brief overview of the modeling approachwe will take; details follow in subsequent subsections. As a first step, we model the underlying trends in physical activity captured via the accelerometer recordings by a non-homogeneous Poisson model. The intensity function, which parametrizes this model, is estimated using functional principal component analysis techniques tailored to our particular context. As we will see later, the resulting intensity function looks similar to the smoothed activity curve in Figure 4.1. It is clear from the graph that we are still missing a significant amount of information that is reflected in the much larger variation in the actual data than what is captured with just the non-homogeneous Poisson model. This motivates us to develop a formal hypothesis test for over-dispersion. As the test confirms our suspicion that a simple non-homogeneous Poisson process is insufficient, we propose two approaches to modify our original model in order to account for over-dispersion.

4.2.1 Estimate Individual Intensity Curve by Multilevel Functional PCA

Under our statistical framework, we model the occurrence of accelerations with a non-homogeneous Poisson process allowing each individual to have his/her own personalized intensity curve. Specifically, if we denote the accelerometer count at time t on the jth day for subject

i as $X_{ij}(t)$, then $X_{ij}(t)$ follows a non-homogeneous Poisson process parametrized by subject i's personal specific intensity function denoted as $\Lambda_i(t)$. To estimate the intensity curve $\Lambda_i(t)$ for each subject using the high-dimensional time-series data obtained via our accelerometer devices, we use principal component methods to reduce dimension. Also, our studies collected accelerometer measures on multiple days for each participant. To account for the hierarchical nature of these data (minutes nested within days within subjects), we will use multilevel functional mixed model methods. Specifically, we propose to use a two-level functional principal component analysis model [DCCP09]. Although this method is published, for the sake of completeness and to introduce our notation, we briefly describe the mathematical underpinnings of this approach below.

As a quick review of functional PCA technique, let $X(t), t \in [0,1]$, be a squared integrable random function with mean $u(t) = E\{X(t)\}$ and covariance function $K(s,t) = cov\{X(s),X(t)\}$. Mercer's theorem gives the following spectral decomposition of $K(s,t)$:

$$K(s,t) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(s)\varphi_k(t)$$

where $\lambda_1 \geq \lambda_2 \geq ...$ are ordered nonnegative eigenvalues and φ_k's are the corresponding orthogonal eigenfunctions with unit L^2 norms. The Karhunen-Loève (KL) expansion of $X(t)$ is $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \varphi_k(t)$, where $\xi_k = \int_0^1 \{X(t) - \mu(t)\}\varphi_k(t)dt$ are uncorrelated random variables with mean zero and variance λ_k. These random variables are called principal component scores.

Now consider the one-way functional ANOVA model:

$$X_{ij}(t) = \mu(t) + \Lambda_i(t) + W_{ij}(t)$$

where $X_{ij}(t)$ is the accelerometer count function measured at time t on the jth day for subject i. $\mu(t)$ represents the overall population mean acceleration function at time t. $\Lambda_i(t)$ is the subject-specific deviation from the overall mean function. $W_{ij}(t)$ is the residual subject- and day-specific

deviation from the subject-mean function. Note, $\Lambda_i(t)$ represents the subject level activity curve that will enable us to characterize subject-to-subject variation in activity. $\mu(t)$ is treated as fixed functions while $\Lambda_i(t)$ and $W_{ij}(t)$ are treated as mean 0 stochastic processes. We call $\Lambda_i(t)$ level 1 functions, and $W_{ij}(t)$ level 2 functions. This model specifies the 2-level hierarchical structure.

Next to account for the minute-level high dimensional data, we will further decompose level 1 and level 2 functions through the Karhunen-Loeve expansion [RS05]:

$$\Lambda_i(t) = \sum_k \xi_{ik}\varphi_k^{(1)}(t)$$

$$W_{ij}(t) = \sum_m \beta_{ijm}\varphi_m^{(2)}(t)$$

where ξ_{ik} and β_{ijm} are level 1 and level 2 principal component scores, and $\varphi_k^{(1)}(t)$, and $\varphi_m^{(2)}(t)$ are level 1 and level 2 eigen-functions. Substitute these into our original model:

$$X_{ij}(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\varphi_k^{(1)}(t) + \sum_{m=1}^{\infty} \beta_{ijm}\varphi_m^{(2)}(t)$$

where $\mu(t)$, $\varphi_k^{(1)}(t)$ and $\varphi_m^{(2)}(t)$ are fixed functional effects, and ξ_{ik} and β_{ijm} are random variables with mean zero. The principal component scores, ξ_{ik}, can be used to distinguish temporal activity patterns between individuals. Before delving into the details of estimating each element in the above expression, we first list some assumptions of the multi-level functional PCA technique [DCCP09]:

A.1 $E(\xi_{ik}) = 0$, $var(\xi_{ik}) = \lambda_k^{(1)}$, for any i, $k_1 \neq k_2$, $E(\xi_{ik_1}\xi_{ik_2}) = 0$;

A.2 $\{\varphi_k^{(1)}(t) : k = 1, 2, ...\}$ is an orthonormal basis of $L^2[0,1]$;

A.3 $E(\beta_{ijm}) = 0$, $var(\beta_{ijm}) = \lambda_m^{(2)}$, for any i, j, $m_1 \neq m_2$, $E(\beta_{ijm_1}\beta_{ijm_2}) = 0$;

A.4 $\{\varphi_m^{(2)}(t) : m = 1, 2, ...\}$ is an orthonormal basis of $L^2[0,1]$;

A.5 $\{\xi_{ik} : k = 1,2,...\}$ are uncorrelated with $\{\beta_{ijm} : m = 1,2,...\}$.

Under assumptions (A.1)-(A.5), let $K_T(s,t) = cov\{X_{ij}(s), X_{ij}(t)\}$ be the overall covariance function, and $K_B(s,t) = cov\{X_{ij}(s), X_{ik}(t)\}$ be the covariance function between level 2 units within the same level 1 unit. Then $K_T(s,t) = \sum_{k=1}^{\infty} \lambda_k^{(1)} \varphi_k^{(1)}(s)\varphi_k^{(1)}(t) + \sum_{m=1}^{\infty} \lambda_m^{(2)} \varphi_m^{(2)}(s)\varphi_m^{(2)}(t)$ and $K_B(s,t) = \sum_{k=1}^{\infty} \lambda_k^{(1)} \varphi_k^{(1)}(s)\varphi_k^{(1)}(t)$. Let $K_W(s,t) := K_T(s,t) - K_B(s,t)$.
We follow the algorithm below [DCCP09]:

Step 1 obtain $\hat{\mu}(t)$, $\hat{K}_T(s,t)$ and $\hat{K}_B(s,t)$ using the method of moments; set $\hat{K}_W(s,t) = \hat{K}_T(s,t) - \hat{K}_B(s,t)$;

Step 2 use eigenanalysis on $\hat{K}_B(s,t)$ to obtain $\hat{\lambda}_k^{(1)}$, $\hat{\varphi}_k^{(1)}(t)$;

Step 3 use eigenanalysis on $\hat{K}_W(s,t)$ to obtain $\hat{\lambda}_m^{(2)}$, $\hat{\varphi}_m^{(2)}(t)$;

Step 4 estimate principal component scores using the projection method detailed in Di's paper

To give some simple examples on how to estimate mean function and covariance functions. Suppose $X_{ij}(t)$ is measured at a set of grid points, $\{t : t = 1,2,...,T\}$, common for every subject and day. Further assume the total number of subjects is I, and the total number of days for each subject is J. Then $\hat{\mu}(t) = \bar{X}_{\cdot\cdot}(t) = \sum_{i,j} X_{ij}(t)/(IJ)$, $\hat{K}_T(t_s,t_r) = \sum_{i,j}\{X_{ij}(t_s) - \hat{\mu}(t_s)\}\{X_{ij}(t_r) - \hat{\mu}(t_r)\}/(IJ)$, and $\hat{K}_B(t_s,t_r) = \sum_i \sum_{j_1 \neq j_2}\{X_{ij_1}(t_s) - \hat{\mu}(t_s)\}\{X_{ij_2}(t_r) - \hat{\mu}(t_r)\}/\{IJ(J-1)\}$.

4.2.2 Account for Over-dispersion via Two Statistical Models

The first part of our model yields personalized intensity functions that capture the most important modes of variation in each subject's physical activity profile. Under our basic framework, the occurrence of acceleration follows a non-homogeneous Poisson process parametrized by these intensity functions. We further develop our model in order to account for the extremely large variation in our data as can be seen in a typical daily profile in Figure 4.1.

Test for Over-dispersion

We first develop a formal statistical test for over-dispersion. The challenge is that our model assumes a non-homogeneous Poisson process, allowing the intensity parameter to vary over time. To conduct hypothesis testing, we would like to fix the intensity parameter to a constant value. To this end, we use the fact that under the definition of non-homogeneous Poisson model, the total number of accelerations that occur between any two time points s and r follows a regular Poisson distribution with intensity $\int_s^r \Lambda(t)dt$. After we obtain the individual intensity function $\Lambda_i(t)$, we divide each daily profile into bins of varying time lengths such that the area under the intensity curve inside each bin is equal to the same constant value Λ. For example, suppose we set the constant $\Lambda = 2000$, and the intensity function for the ith subject is a simple step function:

$$\Lambda_i(t) = \begin{cases} 0 & \text{if } 0 \leq t < 500 \text{ min} \\ 20 & \text{if } 500 \leq t < 650 \text{ min} \\ 10 & \text{if } 650 \leq t < 1000 \text{ min} \\ 15 & \text{if } 1000 \leq t < 1170 \text{ min} \\ 19 & \text{if } 1170 \leq t \leq 1220 \text{ min} \\ 0 & \text{if } 1220 \leq t \leq 1440 \text{ min} \end{cases}$$

Note the intensity function is estimated on minute level (each day has 1440 minutes in total). The 0 value from midnight to 8:20am and from 8:20pm to midnight implies non-wear and hence possible sleep schedule. Under the hypothetical setting, the first bin of this day would be from time 500 min to 600 min since $\int_{500}^{600} 20 = 2000 = \Lambda$. The second bin of this day is from 600 min to 750 min; the third bin of this day is from 750 min to 950 min; the fourth bin of this day is from 950 min to 1100 min; the fifth bin of this day is from 1100 min to 1220 min. Note here that the constant Λ applies to all days and all subjects. Since the intensity functions $\Lambda_i(t)$ are

subject specific, if there are multiple daily profiles for one participant, the binning for every day of this person would be the same as $\Lambda_i(t)$ is the same across all days for subject i. The binning for different subjects, however, could be very different since every person has a different intensity function (detailed in Section 4.2.1).

Once we set our bins as such, under the assumption of non-homogeneous Poisson process, the total number of accelerations that occur within each bin (in our case, the sum of minute level accelerometer counts that fall inside each bin) follows a simple Poisson distribution with rate Λ. In the aforementioned toy example where Λ is set to 2000 and $\Lambda_i(t)$ is a simple step function, the activity sums from each bin on the jth day of this subject are $\{\sum_{t=500}^{t=600} X_{ij}(t), \sum_{t=600}^{t=750} X_{ij}(t), \sum_{t=750}^{t=950} X_{ij}(t), \sum_{t=950}^{t=1100} X_{ij}(t), \sum_{t=1100}^{t=1220} X_{ij}(t)\}$ (where $X_{ij}(t)$ is the accelerometer count measured at time t on this day) and they follow a simple Poisson distribution with rate 2000. Even though the binning for the same subject is the same across different days, the activity sums that fall in the mth bin on different days of the same subject could vary as we observe different activity counts in that bin on different days. In this way we reduce our situation from a non-homogeneous Poisson process to a simple Poisson form. Suppose our algorithm results in a total of M bins across all daily profiles and all subjects, we perform hypothesis testing on the sum of activity counts in each bin, denoted by $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$, using the Central Limit Theorem with 1st order Edgeworth correction. Under the null hypothesis, $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$ are i.i.d $\sim$ Poisson(Λ).

First we list some basic facts about simple Poisson process:

$$E(N_m) = \Lambda$$

$$E(N_m - \Lambda)^2 = \Lambda$$

$$E(N_m - \Lambda)^4 = \Lambda + 3 \times \Lambda^2$$

$$E(N_m - \Lambda)^6 = \Lambda + 25 \times \Lambda^2 + 15 \times \Lambda^3$$

Looking at the second moment, we know $E(N_m-\Lambda)^2=\Lambda$, $var(N_m-\Lambda)^2=E(N_m-\Lambda)^4-E^2(N_m-\Lambda)^2=\Lambda+3\Lambda^2-\Lambda^2=\Lambda+2\Lambda^2$. Let $Y_m=\frac{(N_m-\Lambda)^2-\Lambda}{\sqrt{\Lambda+2\Lambda^2}}$, our simple calculation shows that under the Central Limit Theorem, the quantity $\frac{1}{\sqrt{M}}\sum_{m=1}^{M}Y_m$ converges to standard normal distribution as $M\to\infty$. With the first order Edgeworth correction, the cumulative distribution of of $\frac{1}{\sqrt{M}}\sum_{m=1}^{M}Y_m$ is modified to

$$G(x)=\Phi(x)-\phi(x)(\frac{\gamma(x^2-1)}{6\sqrt{M}})$$

where $\gamma=E(Y_m^3)=\frac{(\Lambda+22\Lambda^2+8\Lambda^3)}{(\Lambda+2\Lambda^2)^{\frac{3}{2}}}$, $\Phi(x)$ is the cumulative distribution function of the standard normal distribution, and $\phi(x)$ is the probability density function of the standard normal distribution. In a later section, we will demonstrate through simulations that the convergence rate using the Central Limit Theorem alone is unsatisfactorily low under our Poisson model with large skew. The Edgeworth expansion yields much more accurate p-value for hypothesis testing.

Two Models to Account for Over-dispersion

Once we are able to provide evidence for over-dispersion, we will improve upon the original non-homogeneous Poisson model to account for it. We propose the following two models:

Model 1 We allow the magnitude of each occurrence of acceleration to vary. This results in a non-homogeneous compound Poisson process. Specifically, given the intensity function $\Lambda_i(t)$ for subject i on the jth day, activity count at time r (min) follows a compound Poisson distribution:

$$\text{activity count at time r, } X_{ij}(r)=\sum_{s=1}^{S_i}V_s$$

where $S_i\sim Poisson\left(\int_{r-1}^{r}\Lambda_i(t)dt\right)$, and V_s's represent the magnitudes of each acceleration. We note that it is theoretically possible to estimate the distribution of V_s non-parametrically. To do this, we again use the sum of accelerometer counts within each bin (as described

in Section 4.2.2) $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1, 2, ..., M\}$. Under our current model, N_m follows a compound Poisson distribution with constant Poisson rate Λ, i.e. $N_m = \sum_{s=1}^{S} V_s$ where $S \sim Poisson(\Lambda)$, and we are trying to estimate the distribution of the discrete random variable V_s. Since N_m is a discrete random variable as well, we first obtain its empirical distribution, i.e. $P(N_m = k)$ for $\forall k$. We then empirically estimate its characteristic function $\phi_{N_m}(t) = \sum_{k=-\infty}^{\infty} P(N_m = k) e^{ikt}$. Next we estimate the characteristic function of the marks V_s through the following equality:

$$\phi_{N_m}(t) = e^{\Lambda(\phi_{V_s}(t)-1)}$$

Once we obtain $\phi_{V_s}(t)$, we can estimate the distribution of V_s using the fact that for discrete random variable, $P(V_s = k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} \phi_{V_s}(t) dt$. This gives the hypothetical outline for evaluating the empirical distribution of the marks V_s. However, the use of characteristic function, especially in the presence of relatively high observed values of N_m, yields very poor consistency. Therefore, we decide to take a parametric approach to model the distribution of V_s. We assume that the magnitude of acceleration V_s follows a negative binomial distribution because of the flexibility of this model and its ability to accommodate high skewness. Furthermore, we have the option to either set the parameters of the negative binomial distribution constant among all subjects or allow the parameters to vary on individual level.

Model 2 We use a non-homogeneous quasi-Poisson process. Instead of regular Poisson distribution where variance must be the same as expectation, we allow variance to change linearly with expectation. In a simple quasi-Poisson setting, if $E(X) = \lambda$, then $var(X) = \theta\lambda$. Here we use θ as the scale parameter to adjust for over-dispersion. Again, we have the option to either set the parameter θ constant among all subjects or allow it to vary on individual level.

Estimation of Over-dispersion Parameter

We will now explain in detail how we estimate the parameter θ in Model 2 as described in the previous section. This will be followed by a brief discussion on the estimation of the parameters for the negative binomial distribution in Model 1. We use the following three methods to estimate θ. Note for each approach, there are two quantities we can use for this estimation: the bin sums $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$ (as described in Section 4.2.2) or the raw data $\{X_{ij}(t) : \forall i, j, t\}$.

Method 1 Set θ constant across all subjects.

$\{N_m\}$: Under the quasi-Poisson model, $E(N_m) = \Lambda$ and $var(N_m) = \theta\Lambda$, therefore $\hat{\theta} = \frac{var(N_m)}{\Lambda}$.

$\{X_{ij}(t)\}$: Under the quasi-Poisson model, $E(X_{ij}(t)) = \mu(t) + \Lambda_i(t)$ and $Var(X_{ij}(t)) = \theta(\mu(t) + \Lambda_i(t))$. Here $\mu(t)$ and $\Lambda_i(t)$ come from the functional PCA model introduced in Section 4.2.1. Simple transformation gives us $E(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}})^2 = \theta$. Suppose we have R daily accelerometer records in total and each day has T time points, we estimate θ using method of moment $\hat{\theta} = \frac{1}{RT}\sum_{ijt}(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}})^2$.

Method 2 Estimate subject specific θ_i for each subject i using method of moments.

$\{N_m\}$: We simply use only the bin sums for subject i, N_m^i, to estimate $\hat{\theta}_i = \frac{var(N_m^i)}{\Lambda}$.

$\{X_{ij}(t)\}$: Similarly, we use only the raw data $X_{ij}(t)$ for subject i to estimate θ_i. Suppose we have J_i daily accelerometer records for subject i and each day has T time points, we set $\hat{\theta}_i = \frac{1}{J_i T}\sum_{jt}(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}})^2$.

Method 3 Estimate subject specific θ_i for each subject i using linear mixed model.

$\{N_m\}$: We line up the bin sums for each daily record of all subjects and perform linear mixed model:

$$\frac{(N_m^{ijb} - \Lambda)^2}{\Lambda} = \alpha_i + \varepsilon_{ijb}$$

where N_m^{ijb} is the bth bin on the jth day for subject i, α_i is the random intercept for subject i, and ε_{ijb} is the random error for the bth bin on the jth day of the ith subject.

$\{X_{ij}(t)\}$: Similarly, we perform linear mixed model on raw data

$$\frac{(X_{ij}(t)-\mu(t)-\Lambda_i(t))^2}{\mu(t)+\Lambda_i(t)} = \alpha_i + \varepsilon_{ijt}$$

where α_i is the random intercept for subject i and ε_{ijt} is the random error for time t on the jth day of the ith subject.

Theorem 1. *The estimators in Method 1 and 2 are (a) consistent; (b) unbiased; and (c) have variances that could be expressed with the first four moments of the variables $\{N_m\}$ and $\{X_{ij}(t)\}$.*

Proof

(a) Given the properties of the non-homogeneous Poisson process, N_m are identically and independently distributed with a quasi-Poisson distribution. The estimators in Method 1 and Method 2 using the bin sums are therefore consistent. Similarly by the definition of non-homogeneous Poisson process, conditional on $\mu(t)$ and $\Lambda_i(t)$, $\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}}$ are identically and independently distributed with $E(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}}) = 0$ and $var(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}}) = \theta$ (under Method 1) or θ_i (under Method 2). Under such conditions, the method of moments estimators in Method 1 and Method 2 are guaranteed to be consistent.

(b) These estimators are also unbiased as $E(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}})^2$
$= var(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}}) + E^2(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}}) = var(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}}) = \theta$ (under Method 1) or θ_i (under Method 2).

(c) Furthermore, we could easily compute the variance of $\hat{\theta}$ coming from the bin sums $\{N_m\}$ in Method 1: $var(\hat{\theta}) = var(\frac{1}{M}\sum_m \frac{(N_m-\Lambda)^2}{\Lambda}) = \frac{var(N_m-\Lambda)^2}{M\Lambda^2} = \frac{var(N_m^2-2\Lambda N_m)}{M\Lambda^2} =$
$\frac{var(N_m^2)+4\Lambda^2 var(N_m)+cov(N_m^2,2\Lambda N_m))}{M\Lambda^2} = \frac{E(N_m^4)-E^2(N_m^2)+4\Lambda^2 var(N_m)+2\Lambda E(N_m^3)-2\Lambda E(N_m^2)E(N_m)}{M\Lambda^2}$ by using

the first four moments of the distribution of N_m. We could also compute the variance of $\hat{\theta}$ coming from raw data $\{X_{ij}(t)\}$ in Method 1: $var(\hat{\theta}) = var(\frac{1}{RT}\sum_{ijt}(\frac{X_{ij}(t)-\mu(t)-\Lambda_i(t)}{\sqrt{\mu(t)+\Lambda_i(t)}})^2)$ $= \sum_{ijt}\frac{E(X_{ij}(t)^4)-E^2(X_{ij}(t)^2)+4(\mu(t)+\Lambda_i(t))^2 var(X_{ij}(t))+2(\mu(t)+\Lambda_i(t))(E(X_{ij}(t)^3)-E(X_{ij}(t)^2)E(X_{ij}(t)))}{R^2T^2(\mu(t)+\Lambda_i(t))^2}$ by using the first four moments of the distribution of $X_{ij}(t)$. These variance computations could easily extend to the estimators in Method 2.

The calculations above provide alternative methods for estimating the over-dispersion parameter. In a later section, we also develop metrics to evaluate and compare the performance of all three methods.

Now we take a look at the compound Poisson model (Model 1 in Section 4.2.2). We focus on the bin sums $\{N_m = \sum_{t^{ij}\in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$ for this discussion. Under this model, the sum of activity counts in each bin follows a compound Poisson process with constant Poisson rate Λ, i.e., $N_m = \sum_{i=1}^{n} V_i$, where $n \sim Poisson(\Lambda)$, and V_i follows a negative binomial distribution. If we denote the expected value of V_i as a and the variance of V_i as b^2, simple calculation shows $E(N_m) = \Lambda a$ and $Var(N_m) = \Lambda(a^2+b^2)$. Under the assumption A.3 of functional PCA in Section 4.2.1 , $E(X_{ij}(t)) = \mu(t) + \Lambda_i(t)$. If we pick a random bin, say the mth bin, $E(N_m) \approx E(\int_{m\text{th bin}} X_{ij}(t)) = \int_{m\text{th bin}} E(X_{ij}(t)) = \int_{m\text{th bin}} (\mu(t)+\Lambda_i(t)) \approx \Lambda$, hence the estimated a here should be very close to 1. The variance term is so large that (a^2+b^2) is in the thousands, hence overwhelmingly dominated by the variance term b. This puts us in a similar situation under the quasi-Poisson model (Model 2 in Section 4.2.2) where $E(N_m) = \Lambda$ and $var(N_m) = \theta\Lambda$ with $b^2 \approx \theta$. Hence in our case the two models would produce very similar results. We therefore only discuss the quasi-Poisson model in future sections. However, we stress that the compound Poisson model provides a different but equally useful framework for accelerometry analysis.

4.3 Application to Data

4.3.1 Study Sample Descriptives

The primary objective of the NIH-funded Transdisciplinary Research on Energetics and Cancer (TREC) Center at UCSD (2011-2017) is to enhance knowledge on the role of obesity and energetics on cancer risk [PCH+13]. To this end, two randomized controlled weight-loss interventions were conducted, the MENU trial (N = 245) in overweight otherwise healthy women, and the Reach for Health trial (N = 333) in postmenopausal overweight breast cancer survivors. In both trials, physical activity was assessed via 7-day accelerometry. We used baseline data from these trials in the current project. Details on the study design and protocols are given below.

The MENU trial was a 12 month behavioral intervention study among 245 overweight non-diabetic women to investigate the role of dietary macronutrient composition on weight loss [LFN+16]. Participants were randomized to one of three diets: a lower fat (20% of energy) and higher carbohydrate (65% of energy) diet; a lower carbohydrate (45% energy) and higher monounsaturated fat (35% energy) diet; or a walnut-rich (35% fat) and lower-carbohydrate (45%) diet. The RfH Study was a 2*2 randomized trial of 333 overweight, postmenopausal early-stage breast cancer survivors, aiming to test the impact of metformin treatment alone, a lifestyle-based intervention alone, both or neither on weight-loss and biomarkers associated with cancer risk [PMN+16].

In both the MENU and RfH trials, height and weight were measured at clinic visits and used to calculate BMI. Physical activity and sedentary behavior were obtained via a triaxial accelerometer, the GT3X Actigraph monitors (ActiGraph, LLC; Pensacola, FL), which is set to collect data at 30 Hz [Bas12a]. The ActiLife program applied a band-pass filter to remove non-human acceleration signal from data and then summarized the signal to counts per minute using a proprietary algorithm [Bas12a]. Information on demographics, lifestyle (e.g., smoking), cancer characteristics and treatment (RfH only), and health status was also collected. Fasting

blood samples were drawn at clinic visits and used to assay glucoregulatory and inflammatory markers.

Women in the MENU Study were on average 50 years old (SD = 10) at study entry, age range 22-72 years. Mean (SD) years of education was 15.6(2.2). Race/ethnicity was self-reported as white non-Hispanic 74.3%, African-American 5.7%, Hispanic 17.1%, Asian 1.6%, Native American 0.4%, and mixed/other race 0.8%. Women in the RfH study were on average 63 years old (SD = 6.9) at study entry, with a mean 2.7 (SD = 2) years from their breast cancer diagnosis; 51% had a college degree; the majority race was White (83%), with 4% African-American, 2% Asian, and 11% of mixed/other race; 11% reported Hispanic ethnicity. Further details on the cohorts are provided in Table 4.1.

4.3.2 Data Preparation

Participants in our trials were directed to wear the accelerometer during waking hours, except while showering. To account for varying start and stop wear times and to maximize information, we retained days with more than 12 hours of consecutive wear time, and only used these "complete profiles" in our analysis. The set of complete profiles consisted of 570 participants, out of the 613 subjects we originally recruited so that $< 8\%$ of the subject sample was discarded. We ended with a total of 3413 days out of 4704 days of records we originally collected, with the number of days per participant ranging from 1 to 14. Lastly, we aligned all daily records by the first minute of device wear, thus allowing each participant to start their daily activity at different times of the day. On average, the participants started wearing the accelerometers at about 7:30am (SD 85min). Most people started their day between 6am and 9am. Furthermore our analysis only used the first 720 points (i.e., 12 hours) of each record. This resulted in a fixed set of grid points $\{t : t = 1,2,...,T\}$, where $T = 720$, for each daily profile. We lost about 137 min per daily record due to this truncation.

4.3.3 Extracting Principal Components

Once the data was ready, we performed functional principal component analysis. An important step was to decide how many principal components to keep in the model. Selecting the number of principal components involved a trade-off between explaining variation in the data, while reducing artifactual patterns. We wanted to find the optimal balance between under-fitting and over-fitting. If $\lambda_r^{(1)}$ denotes the variance explained by the r-th level 1 principal component, the following criteria were used:

$$N = min\{k : \rho_k^{(1)} \geq P_1, \lambda_k < P_2\}, \text{ where } \rho_k^{(1)} = (\lambda_1^{(1)} + ... + \lambda_k^{(1)})/(\lambda_1^{(1)} + ... + \lambda_T^{(1)})$$

with values for P_1 and P_2 chosen to reflect the trade-off. For instance, with $P_1 = 0.9$, and $P_2 = 1/720$, we would need $N = 87$ level 1 components (eigenfunctions). Interestingly, to explain 50% of the variation (i.e. set $P_1 = 0.5$ instead) we would still need 87 level 1 components. If we were only interested in explaining at least 50% of the level 1 variation and did not require that λ_r be less than 1/720, only the first four level 1 principal components were needed. In this paper, we took the conservative approach by setting $P_1 = 0.9$ and $P_2 = 1/720$, resulting in our use of the first 87 level 1 principal components for estimating the intensity functions.

We examined the first four level 1 eigenfunctions in greater detail. Figure 4.2 illustrates the first four level 1 principal component (PC) functions. As expected, the first level 1 PC curve represented an overall vertical shift of the mean activity curve. This component captured total activity volume, so that a participant with a high score on this component was on average more physically active than one with a lower score. The second level 1 PC curve emphasized variations in the very early parts of the day. The third level 1 PC curve reflected variation contrast between morning activity and activity throughout the rest of the day. The fourth level 1 PC curve focused on the variation in the middle and later parts of the day. Thus although not the focus of the current work, the principal components can be used to explain temopral variation in activity patterns.

4.3.4 Over-dispersion

Test for Over-dispersion

Edgeworth Expansion Simulations To perform the over-dispersion test, we set bin size $\Lambda =$ 2130. This particular value was chosen in order to obtain on average 100 bins for each daily record. We first performed simulations of Poisson distribution with rate 2130 and compared the cumulative probability at critical value of 1.96 between the results from Central Limit Theorem (CLT) and those from Central Limit Theorem with 1st order Edgeworth Correction. Figure 4.3 to 4.5 illustrate such comparisons (with sample size 10, 100, and 1000) of the difference between the true cumulative probability estimated during each simulation and the theoretical cumulative probability values (CLT vs CLT+Edgeworth). Figure 4.3 shows that at a small sample size ($n = 10$) neither CLT nor the Edgeworth expansion did a good job in estimating p-value. But it was clear that CLT alone performed a lot worse than with Edgeworth correction. It seemed that adding the correction term over-compensated the error in using CLT. At larger sample sizes ($n = 100$, $n = 1000$), Edgeworth expansion produced highly accurate estimates for cumulative probabilities whereas CLT converged to the true value at an unsatisfactorily slow rate.

Test for Over-dispersion Having established that Edgeworth expansion results in superior estimation of p-value under our framework, we applied Central Limit Theorem with 1st order Edgeworth correction and performed hypothesis testing on the sum of activity counts in each bin $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1, 2, ..., M\}$ (as described in Section 4.2.2). Our test resulted in a p-value < 0.001 and thus provided strong evidence of over-dispersion.

Over-dispersion Parameter Estimates

We proceeded to estimate the over-dispersion parameter as detailed in Section 4.2.2. Under Method 1, where we assume constant θ across participants, $\hat{\theta} = 4119$ using $\{N_m\}$ and $\hat{\theta} = 4756$ using $\{X_{ij}(t)\}$. In Methods 2 and 3, we estimate a subject-specific dispersion parameter

θ_i, and hence calcualted summary statistics across subjects. Under Method 2, $\hat{\theta}_i$ has an average of 3753, median of 3320, and SD of 1986 using $\{N_m\}$ whereas $\hat{\theta}_i$ has an average of 3047, median of 1805, and SD of 2917 using $\{X_{ij}(t)\}$. Under Method 3, $\hat{\theta}_i$ has an average of 3866, median of 3428, and SD of 1844 using $\{N_m\}$ whereas $\hat{\theta}_i$ has an average of 3098, median of 1889, and SD of 2842 using $\{X_{ij}(t)\}$. Using raw data produced higher estimate in Method 1 but lower estimates in Method 2 and 3. Raw data also resulted in greater variance in estimation. This is expected, as bin summing can be viewed as a form of averaging/smoothing on the original data.

4.3.5 Visualizing True versus Simulated Data

To visualize the ability of our extended model to reproduce the original accelerometer data, we plotted (Figure 4.11) the true accelerometer data, the simulated data with regular Poisson intensity function (estimated via multi-level functional PCA from Section 4.2.1), and the simulated data with quasi-Poisson intensity function (Model 2 Section 4.2.2) for a randomly selected daily profile. Here the variance parameter θ for the quasi-Poisson model was estimated with Method 3 (Section 4.2.2) using bin sums $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$ (Section 4.2.2). It is clear from the graph that the quasi-Poisson model more successfully recreated the large peaks in the activity profile than the regular Poisson model. The simulated data tracks the real data relatively closely.

4.3.6 Health Outcome Analysis

We next demonstrate the utility of our model in public health research. The homeostatic model assessment (HOMA) index is a function of blood glucose and insulinand quantifies beta-cell function. HOMA is a widely used measure of glucoregulation and insulin resistance. In order to test associations between temporal activity patterns and HOMA, we performed multiplelinear regression of HOMA on activity patterns represented by the first four level 1 principal component

scores (Section 4.2.1) as well as the variance parameter θ (Model 2 Section 4.2.2) or the mark distribution parameters (Model 1 Section 4.2.2), adjusted for age, education, body mass index (BMI), smoke indicator, and cancer indicator. More specifically, we fit the following regression models:

$$\text{HOMA}_i = \beta_0 + \beta_1\xi_{i1} + \beta_2\xi_{i2} + \beta_3\xi_{i3} + \beta_4\xi_{i4} + \beta_5\theta_i + \text{other covariates}_i + \varepsilon_i$$

$$\text{HOMA}_i = \beta_0 + \beta_1\xi_{i1} + \beta_2\xi_{i2} + \beta_3\xi_{i3} + \beta_4\xi_{i4} + \beta_5\text{E}_i(V_s) + \beta_6\text{Var}_i(V_s) + \text{other covariates}_i + \varepsilon_i$$

Here ξ_{i1}, ξ_{i2}, ξ_{i3}, and ξ_{i4} refer to the first, second, third, and fourth level 1 PC scores for subject i. θ_i is the variance parameter for subject i estimated using Method 3 (Section 4.2.2) with bin sums $\{N_m = \sum_{t^{ij}\in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$ (Section 4.2.2). $\text{E}_i(V_s)$ and $\text{Var}_i(V_s)$ are the expectation and variance of the marks for subject i estimated using Method 3 (Section 4.2.2) but with raw data $\{X_{ij}(t) : \forall i,j,t\}$. Other covariates include age, education (1 if college and above; 0 otherwise), BMI, smoke (1 if yes; 0 if no), and cancer (1 if yes; 0 if no). Results are shown in Table 4.2.

In the first regression PC1 scores, PC3 scores, and θ were significantly associated with HOMA index. As we discussed before (Figure 4.2), PC1 measured the total volume of physical activity. The negative association between PC1 scores and HOMA implied that higher overall activity level was linked to lower HOMA which was an indication of lower insulin resistance. PC3 reflected the contrast between morning activity and activity throughout the rest of the day. Higher PC3 scores implied less activity in the morning and higher activity later (Figure 4.2) and were related to better health (lower HOMA) as shown in the negative coefficient in the regression result. The variance parameter θ also had a significant negative association with HOMA. To further understand what information θ was capturing, we plotted the activity profile of the person with the largest θ against that of the person with the smallest θ in Figure 4.10. Our rationale for including this quantity in the regression analysis was that people with large θ tended to have

greater and possibly more frequent oscillations in their activity profiles, which are hypothesized to be associated with a more active and healthy lifestyle. We used these θ parameters as a measure of the total variation in a person's daily activity. Figure 4.10 lends support to our theory.

In the second regression, PC1 scores and PC3 scores were again significant predictors for HOMA. Interestingly however, only the expectation of the marks was significant in this model. The variance of the marks, which theoretically should provide very similar measures/information to the θ parameter in the first regression model, was no longer significant here.

4.4 Model Evaluation

In the previous sections (4.2), we have developed an extended modeling framework which accounts for over-dispersion, and proposed several estimators. A logical next step is to compare the performance of these different methods with respect to an appropriate metric. In particular, we want to compare a true daily record against the corresponding simulated daily record under the following models: 1. regular non-homogeneous Poisson process with intensity function estimated via multi-level functional PCA; 2. non-homogeneous quasi-Poisson process (Model 2 Section 4.2.2) allowing the variance to increase linearly with expectation, with the variance parameter θ estimated through the three different methods mentioned in Section 4.2.2. We devised two metrics to systematically evaluate the performance of our models.

4.4.1 Compare Correlations

We obtained the Pearson correlation between each actual daily profile and the simulated daily profile under the different models as described above. Figure 4.6 shows the side-by-side boxplot comparing the distribution of these correlations under the different models. It was clear that the simple Poisson model performed poorly in comparison to the improved models in this metric. Using bin sums $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1,2,...,M\}$ yielded very

consistent performances across three methods. In method 2 and 3, it out-performed using raw data $\{X_{ij}(t) : \forall i, j, t\}$ whereas in method 1 the reverse happened. Among the three methods, to our surprise, method 1 gave slightly superior performance compared to the other two methods which produced subject specific estimates. Based on this result, we concluded that in the case where correlation was of primary interest to research objectives, the simple Method 1 estimator (i.e., constant θ for all subjects) should be used on raw data $\{X_{ij}(t)\}$ to estimate the variance parameter θ.

4.4.2 Compare Frequencies in Preset Intervals

Our second metric captured discrepancies in activity distributions between the true and simulated distributions. For each daily record (true and simulated), we computed the frequencies of activity count values falling within a set of predetermined intervals. We then calculated, for each daily record, the L_2 distance between the set of frequencies of the true data and that of the simulated data (under the different models mentioned above). We use these quantities as a measure of the distance between the true data probability distribution and the simulated data probability distribution for each day. We first chose a set of finely delineated intervals at an increment of 50 from 0 to 2000, i.e. $\{[0,50],[50,100],...,[1950,2000]\}$. We used 2000 as the upper bound because it was extremely rare to observe values above this threshold. Figure 4.7 shows the side-by-side box-plot illustrating the distance measures between the true data and the simulated data under the different models/methods. Again the simple Poisson model clearly showed the worst performance. The improved models could roughly cut the L_2-distance in half. All the models incorporating over-dispersion had similar performance, although, contrary to what we observed under the correlation metrics, using raw data $\{X_{ij}(t) : \forall i, j, t\}$ out-performed using bin sums $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1, 2, ..., M\}$ under method 2 & 3 and vice versa under method 1. In general, performance improved from method 1 to method 3. We concluded that when frequency comparisons were of importance to research goals, we should use linear

mixed model on raw data to estimate the variance parameter θ_i for each subject. Furthermore, Figure 4.8 shows the aggregated histograms under different models. Each step in the histogram represents one of the intervals. For example, the first step represents the interval $[0, 50]$, the second step represents the interval $[50, 100]$, and so on. The value at each step represents the average frequency (per day) of all the true data or simulated data (under different models) occurring in that interval. The quasi-Poisson model in this graph refers to method 3 using raw data. It was apparent here that the data produced by the improved model was much more similar to the true data in distribution compared to the data produced by the regular Poisson model. The regular Poisson model spread the data more evenly amongst all the intervals and failed to capture the extreme skewness in the original data.

Finally, we performed the same analysis on a different set of intervals $\{[0, 100], [100, 1952], [> 1952]\}$. These values are of special interest to public health research since activity counts below 100 are considered sedentary and activity counts above 1952 are moderate and vigorous physical activity (MVPA). The result (in Figure 4.9) showed essentially the same trend as before.

Lastly we performed similar analysis based on the KolmogorovSmirnov distance between the true data and the simulated data under different models/methods. The results were similar to Figure 4.7 and 4.9 and hence we did not present them here.

4.5 Conclusion

We have presented a novel statistical framework to analyze accelerometer data in this paper. We started with a naive non-homogeneous Poisson process to describe the underlying trend of physical activity. This was achieved by applying the multi-level functional principal component analysis technique. We then tested the data for over-dispersion based on sound mathematical theory and improved upon the naive model by accounting for over-dispersion. We proposed two different approaches/models to modify the base model (regular Poisson process) and for

each improved model further developed three different methods to estimate the additional model parameters. In addition, we devised two metrics to evaluate our models. Eventually, we were able to not only gain insight into temporal physical activity patterns but also relate this insight to health outcomes. Conventionally, research using accelerometer data defaults to simple summary sensor statistics such as average activity count per minute, number of minutes in moderate to vigorous physical activities (MVPA) per day, and number of minutes in sedentary time per day, ignoring the full spectrum of activity profiles provided by accelerometer devices. Our model provides an enriched framework to study physical activity.

We recognized some limitations of the model such as the interpretability of the more intricate/complex activity trend beyond the first four level 1 principal component functions. Moreover, we have not incorporated the information provided by the level 2 principal components. Further studies can be conducted to investigate these components.

On the other hand, we saw a lot of potential for further development and application of our model. Little was known about how variation in diurnal activity patterns might impact health. Applying state-of-the-art statistical methods and novel computational tools to unique accelerometer-based physical activity data, we took a first step in addressing this gap. Our model provided useful insights into human physical activity patterns and the implications for human health. Furthermore, this paper only focused on baseline data from two sample studies. For longitudinal studies, we could simply extend our model from a two level functional PCA model into a three level model incorporating phase change as the third level variation. In other words, we can obtain data collected at different phases (baseline, follow-up, etc) and study the dynamic changes of physical activity patterns for each individual subject. This approach could prove to be relevant for most intervention studies. Lastly, the methods we developed were robust against missing accelerometer values and could generalize to other populations. Our work could definitely be applicable to other research areas, such as GPS tracking and spatial epidemiology.

4.6 Acknowledgements

Chapter 4, in full, is a reprint of the material as it is under review in “Using quasi-Poisson processes to model accelerometer data”. Selene Yue Xu, Jacqueline Kerr, Suneeta Godbole, Ruth E Patterson, Cheryl L Rock, Ian Abramson, and Loki Natarajan. The dissertation/thesis author was the primary investigator and author of this paper.

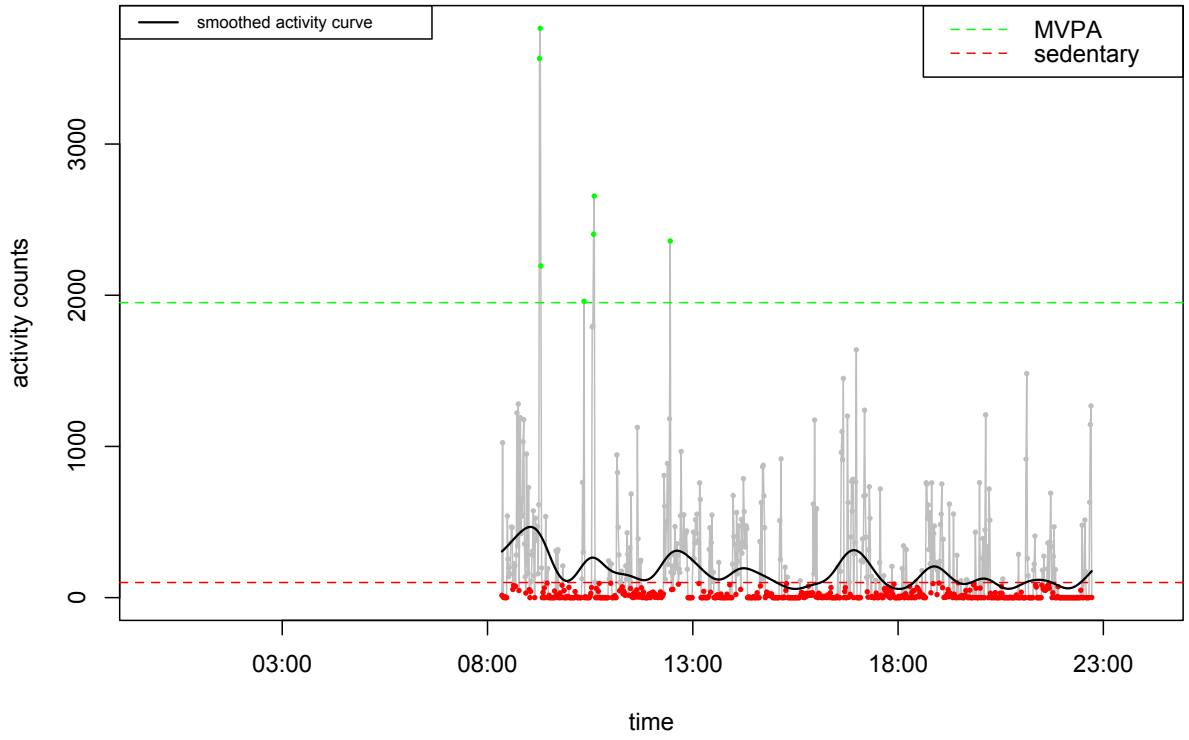

Figure 4.1: A Typical Day of Accelerometer Data

Table 4.1: Characteristics of the two study cohorts (mean(SD) or percentages). Note: *median (1st quartile, 3rd quartile); $^{+}$ IS = insulin sensitive; IR = insulin resistant

	RfH study	MENU study
Age	63.1 (6.9)	50.8 (10.0)
BMI	31.2 (5.0)	34.2 (3.4)
Education	51.7% college and above; 48.3% o.w	56.0% college and above; 44.0% o.w
Smoke	45.2% Yes; 54.7% No	27.6% Yes; 70.0% No; 2.5% NA
Cancer	48% stage 1; 35% stage 2; 17% stage 3	none
HOMA*	2.89 (2.05, 4.14)	3.05 (2.25, 4.64)
Insulin Status^{+}	53.8% IS; 46.2% IR	48.6% IS; 51.4% IR

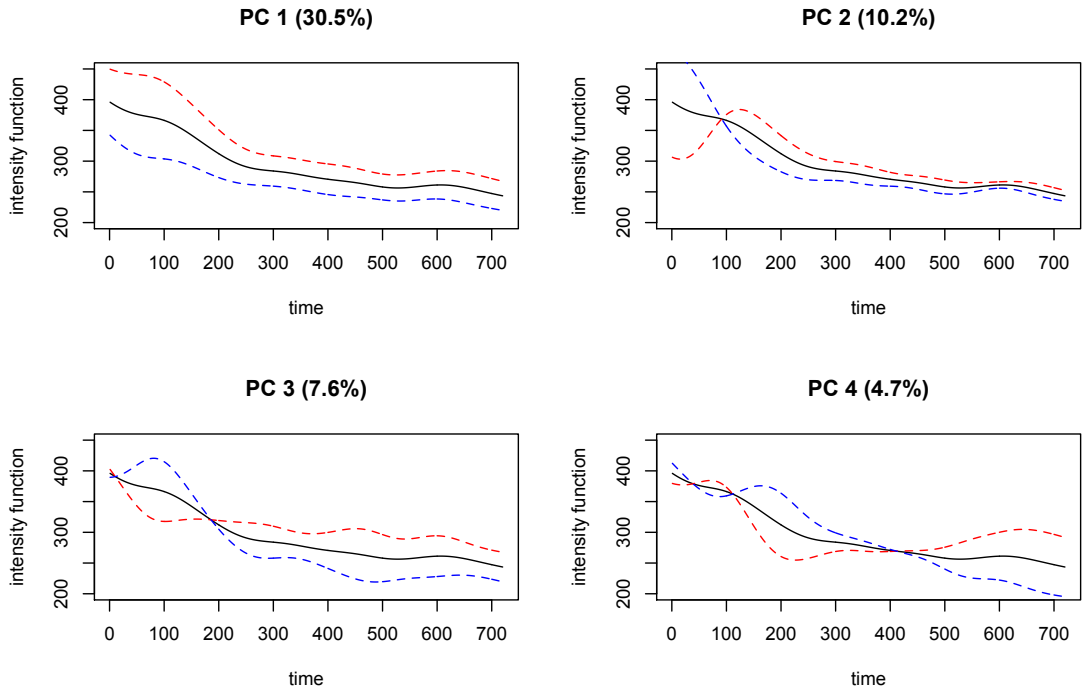

Figure 4.2: (MENU + RfH) The mean intensity curve and the effects of adding (red) and subtracting (blue) a suitable multiple of the first four level-1 PC curves

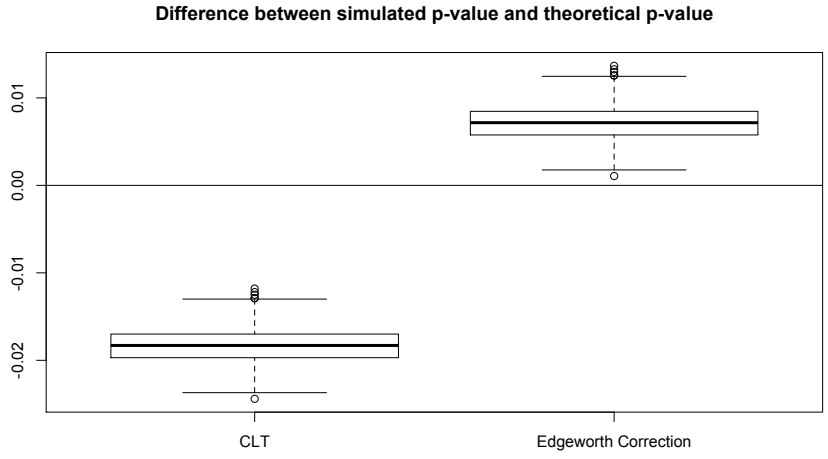

Figure 4.3: The difference between the true cumulative probability estimated during each simulation and the theoretical estimation (CLT vs CLT + Edgeworth) of the cumulative probability at sample size $n = 10$.

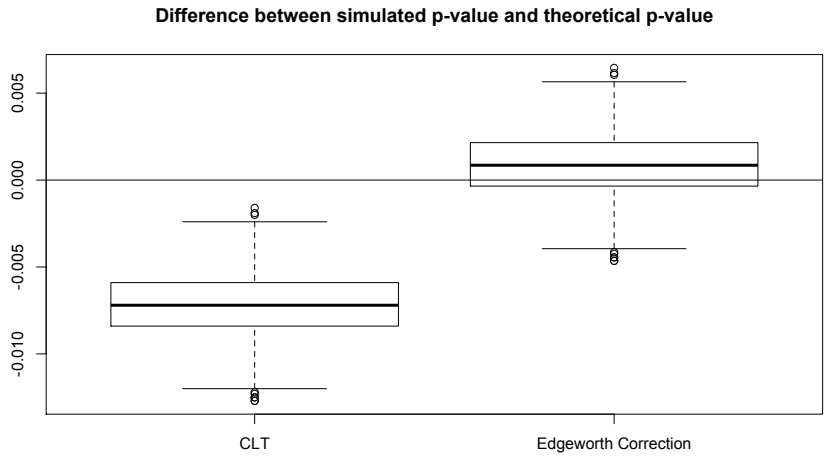

Figure 4.4: The difference between the true cumulative probability estimated during each simulation and the theoretical estimation (CLT vs CLT + Edgeworth) of the cumulative probability at sample size $n = 100$.

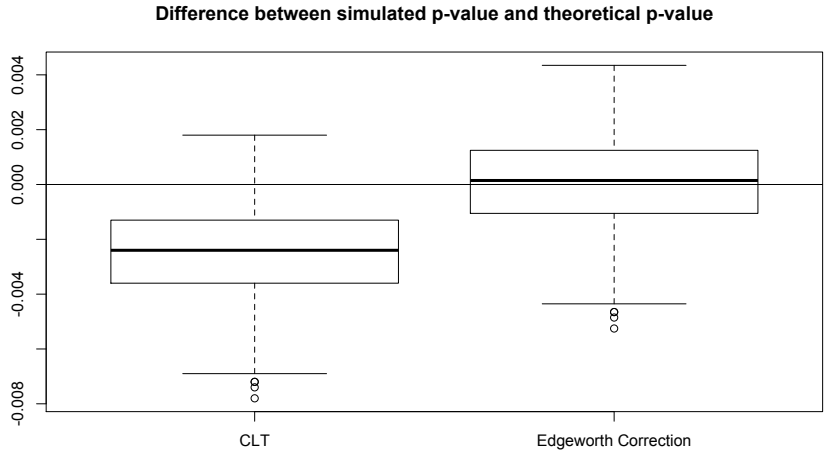

Figure 4.5: The difference between the true cumulative probability estimated during each simulation and the theoretical estimation (CLT vs CLT + Edgeworth) of the cumulative probability at sample size $n = 1000$.

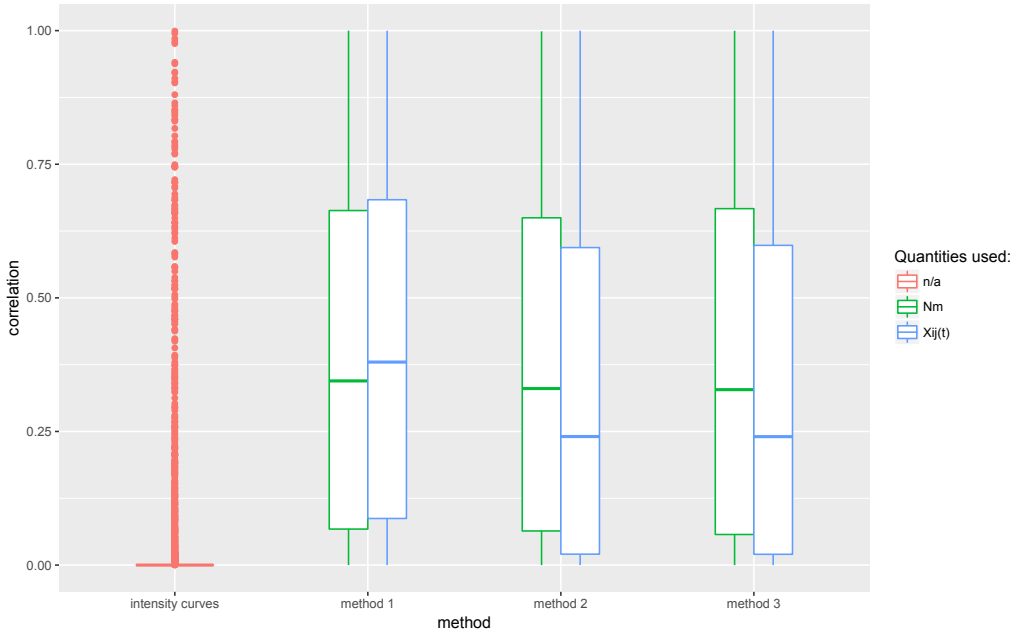

Figure 4.6: Compare Pearson correlation between actual daily profile and the simulated daily profile under different models/methods.

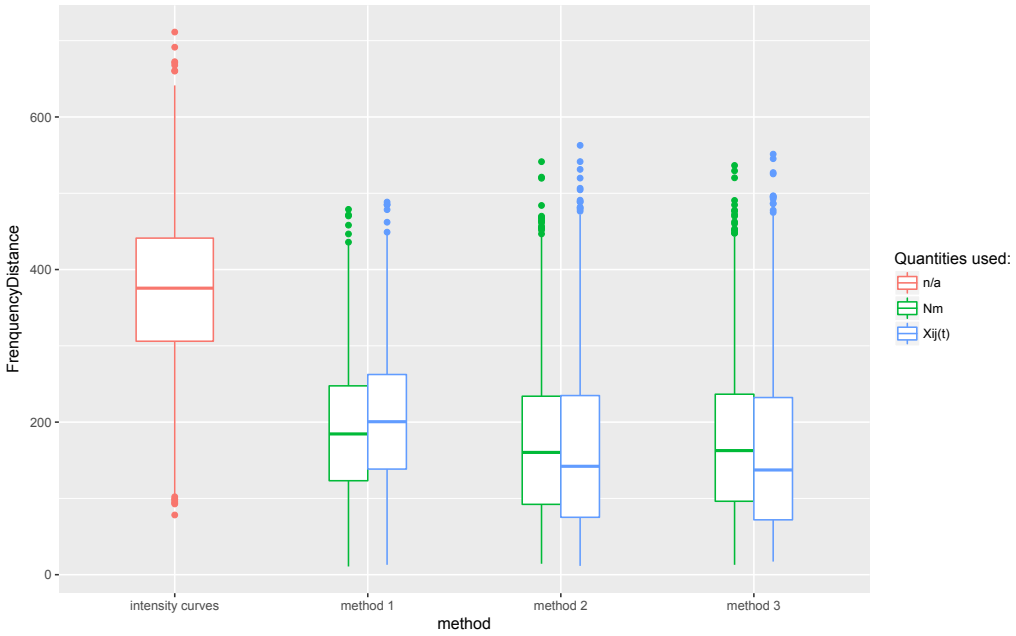

Figure 4.7: Compare the L_2 distance of the frequencies in the preset intervals $\{[0,50],[50,100],...,[1950,2000]\}$ between actual daily profile and the simulated daily profile under different models/methods.

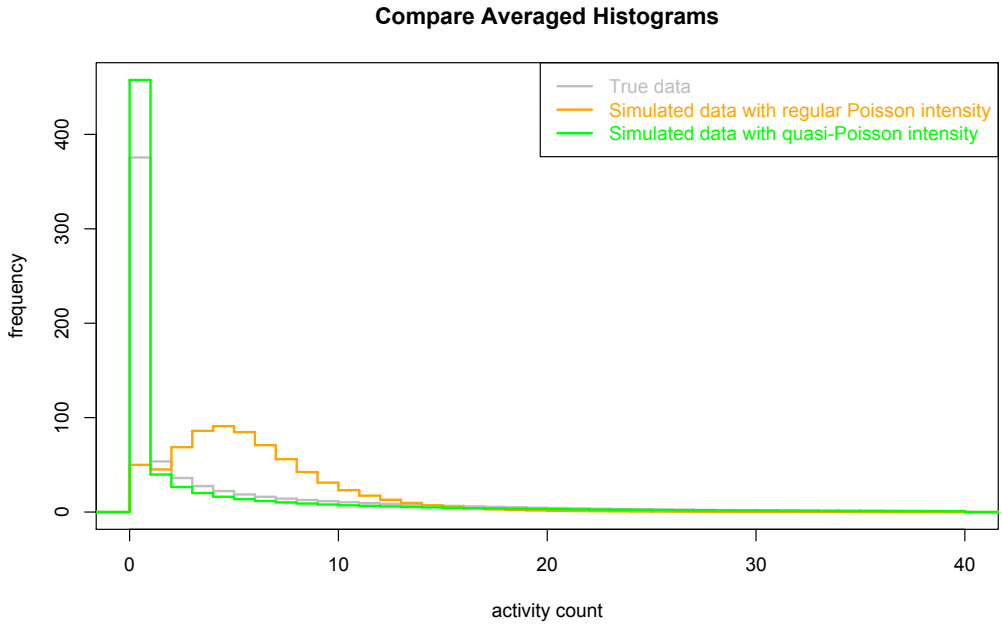

Figure 4.8: Compare the average frequencies in the preset intervals $\{[0,50],[50,100],...,[1950,2000]\}$ between actual daily profile and the simulated daily profile under method 3 using raw data.

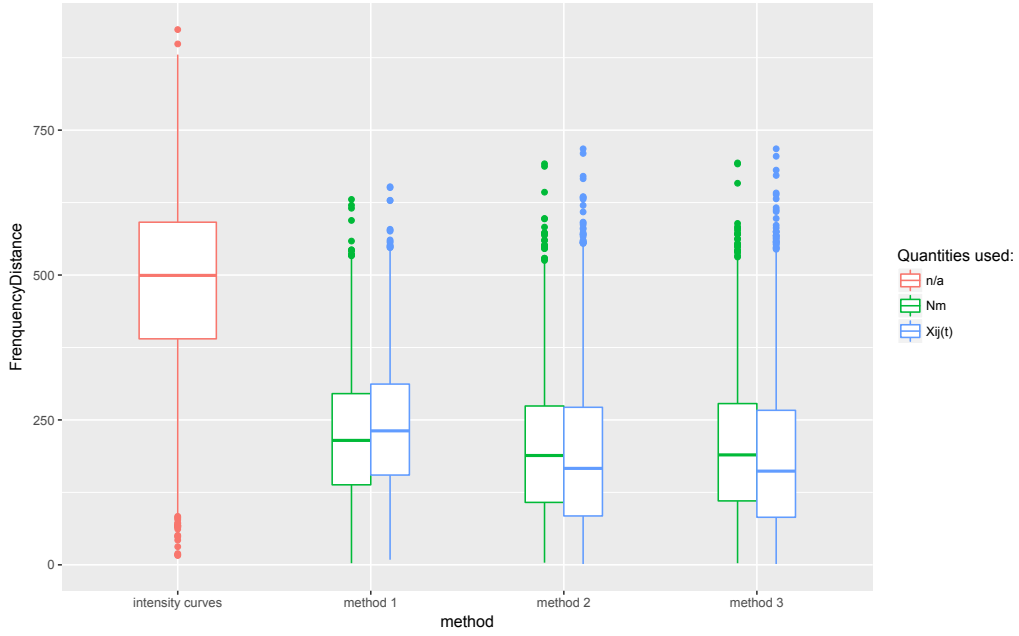

Figure 4.9: Compare the L_2 distance of the frequencies in the preset intervals $\{[0,100],[100,1952],[>1952]\}$ (distinguish among sedentary behavior, regular activity, and moderate to vigorous activity) between actual daily profile and the simulated daily profile under different models/methods.

Table 4.2: Linear regression of HOMA on the first four PC scores and the variance parameter θ / the mark distribution parameters (after adjusted for age, education, smoke, BMI, and cancer status). Here all PC scores, the variance parameter θ, and the mark distribution parameters (expectation and variance of marks) have been standardized to produce comparable results. θ is estimated using Method 3 (Section 4.2.2) with bin sums $\{N_m = \sum_{t^{ij} \in m\text{th bin}} X_{ij}(t) : m = 1, 2, ..., M\}$ (Section 4.2.2). The mark distribution parameters are also estimated using Method 3 (Section 4.2.2) but with raw data $\{X_{ij}(t) : \forall i, j, t\}$.

Dependent Variable	Independent Variables	Coefficient	SE	P-value	Confidence Interval
HOMA	PC1	-0.352	0.101	0.001	(-0.550, -0.153)
	PC2	-0.017	0.093	0.857	(-0.166, 0.199)
	PC3	-0.194	0.094	0.039	(-0.378, -0.010)
	PC4	0.105	0.093	0.257	(-0.077, 0.288)
	θ	-0.220	0.100	0.028	(-0.416, -0.024)
HOMA	PC1	-0.359	0.103	0.001	(-0.562, -0.156)
	PC2	-0.010	0.093	0.915	(-0.193, 0.174)
	PC3	-0.201	0.093	0.032	(-0.384, -0.017)
	PC4	0.103	0.093	0.269	(-0.080, 0.285)
	E(marks)	-0.350	0.120	0.004	(-0.586, -0.114)
	Var(marks)	0.123	0.135	0.363	(-0.142, 0.388)

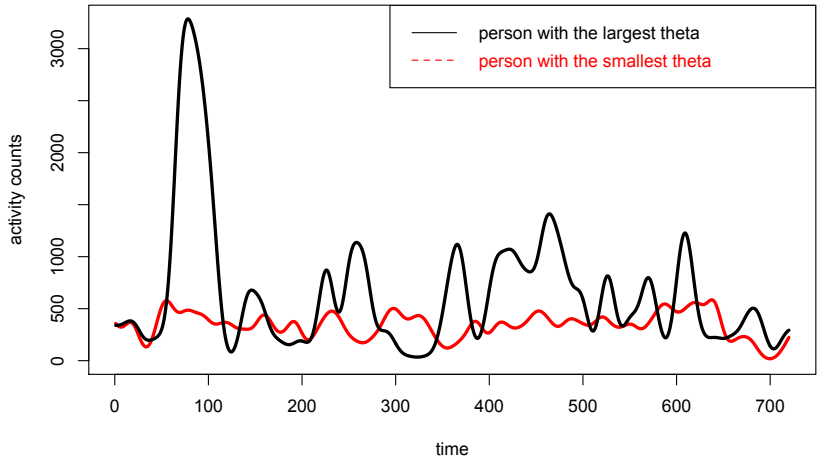

Figure 4.10: Compare a randomly chosen daily profile of the person with the largest variance parameter θ to that of the person with the smallest θ.

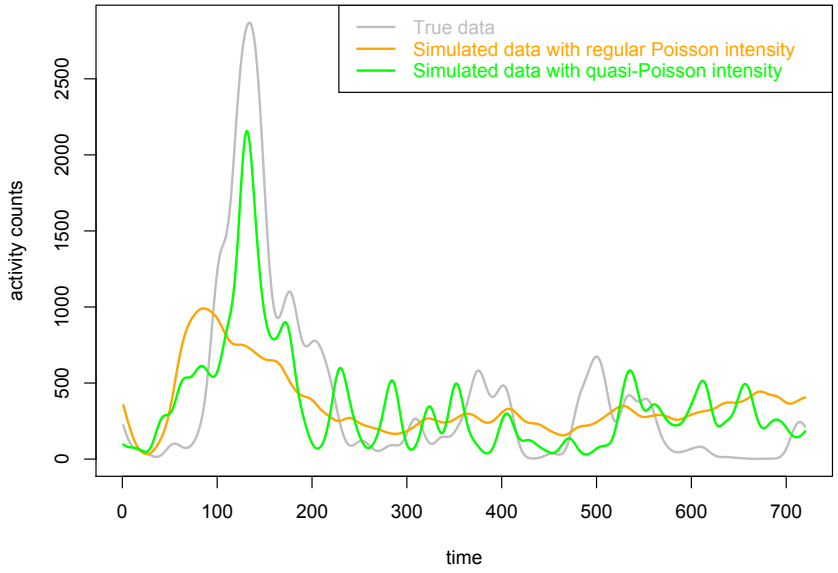

Figure 4.11: True data VS simulated data with regular Poisson intensity function VS simulated data with quasi-Poisson intensity function for a randomly selected daily activity profile.

Bibliography

[AILM+06] Sonia Ancoli-Israel, Lianqi Liu, Matthew R Marler, Barbara A Parker, Vicky Jones, Georgia Robins Sadler, Joel Dimsdale, Mairav Cohen-Zion, and Lavinia Fiorentino. Fatigue, sleep, and circadian rhythms prior to chemotherapy for breast cancer. *Supportive Care in Cancer*, 14(3):201–209, 2006.

[AILR+14] Sonia Ancoli-Israel, Lianqi Liu, Michelle Rissling, Loki Natarajan, Ariel B Neikrug, Barton W Palmer, Paul J Mills, Barbara A Parker, Georgia Robins Sadler, and Jeanne Maglione. Sleep, fatigue, depression, and circadian activity rhythms in women with breast cancer before and after treatment: a 1-year longitudinal study. *Supportive Care in Cancer*, 22(9):2535–2545, 2014.

[ALC+15] Nancy E Avis, Beverly J Levine, L Douglas Case, Elizabeth Z Naftalis, and Kimberly J Van Zee. Trajectories of depressive symptoms following breast cancer diagnosis. *Cancer Epidemiology and Prevention Biomarkers*, 24(11):1789–1795, 2015.

[ARS+16] Charissa Andreotti, James C Root, Sanne B Schagen, Brenna C McDonald, Andrew J Saykin, Thomas M Atkinson, Yuelin Li, and Tim A Ahles. Reliable change in neuropsychological assessment of breast cancer survivors. *Psycho-Oncology*, 25(1):43–50, 2016.

[ATdMR10] Alex Aussem, Andre Tchernof, Sergio Rodrigues de Morais, and Sophie Rome. Analysis of lifestyle and metabolic predictors of visceral obesity with bayesian networks. *BMC bioinformatics*, 11(1):487, 2010.

[Bas12a] David R Bassett. Device-based monitoring in physical activity and public health research. *Physiological measurement*, 33(11):1769, 2012.

[Bas12b] D.R. Bassett. Device-based monitoring in physical activity and public health research. *Physiol. Meas.*, 33:1769–1783, 2012.

[BBB+10] Susan L Beck, Ann M Berger, Andrea M Barsevick, Bob Wong, Katie A Stewart, and William N Dudley. Sleep quality after initial chemotherapy for breast cancer. *Supportive Care in Cancer*, 18(6):679–689, 2010.

[BBBC06] Katherine L Byar, Ann M Berger, Suzanne L Bakken, and Melissa A Cetak. Impact of adjuvant breast cancer chemotherapy on fatigue, other symptoms, and quality of life. In *Oncology nursing forum*, volume 33, page E18. Oncology Nursing Society, 2006.

[BBGAS13] RMDMPH Ballard-Barbash, Stephanie M George, Catherine M Alfano, and KMPHP Schmitz. Physical activity across the cancer continuum. *Oncology*, 27(6):589–592, 2013.

[BCM94] F.C. Bijnen, C.J. Caspersen, and W.L. Mosterd. Physical inactivity as a risk factor for coronary heart disease: a WHO and International Society and Federation of Cardiology position statement. *Bull World Health Organ*, 72:1–4, 1994.

[BHJ+92] John E Brazier, R Harper, NM Jones, A O'cathain, KJ Thomas, T Usherwood, and L Westlake. Validating the sf-36 health survey questionnaire: new outcome measure for primary care. *Bmj*, 305(6846):160–164, 1992.

[BL14] Concha Bielza and Pedro Larranaga. Bayesian networks in neuroscience: a survey. *Frontiers in computational neuroscience*, 8:131, 2014.

[Bla09] S.N. Blair. Physical inactivity: the biggest public health problem of the 21st century. *Br J Sports Med*, 43:1–2, 2009.

[BM08] Ann Malone Berger and Sandra A Mitchell. Modifying cancer-related fatigue by optimizing sleep quality. *Journal of the National Comprehensive Cancer Network*, 6(1):3–13, 2008.

[BOF+15] Aviroop Biswas, Paul I Oh, Guy E Faulkner, Ravi R Bajaj, Michael A Silver, Marc S Mitchell, and David A Alter. Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: a systematic review and meta-analysis. *Annals of internal medicine*, 162(2):123–132, 2015.

[Bow08] Julienne E Bower. Behavioral symptoms in patients with breast cancer and survivors. *Journal of Clinical Oncology*, 26(5):768–777, 2008.

[BRM+89] Daniel J Buysse, Charles F Reynolds, Timothy H Monk, Susan R Berman, and David J Kupfer. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry research*, 28(2):193–213, 1989.

[BSGB98] Ralph HB Benedict, David Schretlen, Lowell Groninger, and Jason Brandt. Hopkins verbal learning test–revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, 12(1):43–55, 1998.

[BYM+10] Daniel J Buysse, Lan Yu, Douglas E Moul, Anne Germain, Angela Stover, Nathan E Dodds, Kelly L Johnston, Melissa A Shablesky-Cade, and Paul A Pilkonis. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep*, 33(6):781–792, 2010.

[CCT10] R. Colley, Gorber S. Connor, and M.S. Tremblay. Quality control and data reduction procedures for accelerometry-derived measures of physical activity. *Health Rep*, 21:63–9, 2010.

[CGJ+11] R.C. Colley, D. Garriguet, I. Janssen, et al. Physical activity of Canadian adults: accelerometer results from the 2007 to 2009 Canadian Health Measures Survey. *Health Rep*, 22:7–14, 2011.

[CHM+05] D.J. Catellier, P.J. Hannan, D.M. Murray, et al. Imputation of missing data when measuring physical activity by accelerometry. *Med Sci Sports Exerc*, 37:S555–62, 2005.

[CIMS15] Judith E Carroll, Michael R Irwin, Sharon Stein Merkin, and Teresa E Seeman. Sleep and multisystem biological risk: a population-based study. *PloS one*, 10(2):e0118467, 2015.

[CLMB11] L1 Choi, Z. Liu, C.E. Matthews, and M.S. Buchowski. Validation of accelerometer wear and non-wear time classification algorithm. *Med Sci Sports Exerc.*, 43:357–64, 2011.

[CMT+13] Barbara Collins, Joyce MacKenzie, Giorgio A Tasca, Carole Scherling, and Andra Smith. Cognitive effects of chemotherapy in breast cancer patients: a dose–response study. *Psycho-Oncology*, 22(7):1517–1527, 2013.

[CRWTT03] Eugenia E Calle, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of us adults. *New England Journal of Medicine*, 348(17):1625–1638, 2003.

[CSSF06] Ester Cerin, Brian E Saelens, James F Sallis, and Lawrence D Frank. Neighborhood environment walkability scale: validity and development of a short form. *Medicine & Science in Sports & Exercise*, 38(9):1682–1691, 2006.

[CWC15] Amanda J Cox, Nicholas P West, and Allan W Cripps. Obesity, inflammation, and the gut microbiota. *The lancet Diabetes & endocrinology*, 3(3):207–215, 2015.

[DCCP09] Chong-Zhi Di, Ciprian M Crainiceanu, Brian S Caffo, and Naresh M Punjabi. Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458, 2009.

[DHLZ13] P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Oxford Univ. Press, 2013.

[DRH+17] Marie-Rose Dwek, Lorna Rixon, Catherine Hurt, Alice Simon, and Stanton Newman. Is there a relationship between objectively measured cognitive changes in patients with solid tumours undergoing chemotherapy treatment and their health-related quality of life outcomes? a systematic review. *Psycho-oncology*, 26(10):1422–1432, 2017.

[EKG+15] K. Ellis, J. Kerr, S. Godbole, J. Staudenmayer, and G. Lanckriet. Hip and wrist accelerometer algorithms for free-living behavior classification. *Med Sci Sports Exerc.*, 2015.

[ESV+15] Sean M Ewings, Sujit K Sahu, John J Valletta, Christopher D Byrne, and Andrew J Chipperfield. A bayesian network for modelling blood glucose concentration and exercise in type 1 diabetes. *Statistical methods in medical research*, 24(3):342–372, 2015.

[FBB+11] Francesca Foltran, Paola Berchialla, Riccardo Bigi, Giuseppe Migliaretti, Alberto Bestetti, and Dario Gregori. Understanding coronary atherosclerosis in relation to obesity: is getting the distribution of body fatness using dual-energy x-ray absorptiometry worth the effort? a novel perspective using bayesian networks. *Journal of evaluation in clinical practice*, 17(1):32–39, 2011.

[FLNP00] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[FPTBV+16] Pilar Fuster-Parra, P Tauler, M Bennasar-Veny, A Ligeza, AA Lopez-Gonzalez, and A Aguilo. Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. *Computer methods and programs in biomedicine*, 126:128–142, 2016.

[Fra15a] Richard B Francoeur. Using an innovative multiple regression procedure in a cancer population (part 1): detecting and probing relationships of common interacting symptoms (pain, fatigue/weakness, sleep problems) as a strategy to discover influential symptom pairs and clusters. *OncoTargets and therapy*, 8:45, 2015.

[Fra15b] Richard B Francoeur. Using an innovative multiple regression procedure in a cancer population (part ii): fever, depressive affect, and mobility problems clarify an influential symptom pair (pain–fatigue/weakness) and cluster (pain–fatigue/weakness–sleep problems). *OncoTargets and therapy*, 8:57, 2015.

[GACT+17] Margaret E Gruen, Marcela Alfaro-Córdoba, Andrea E Thomson, Alicia C Worth, Ana-Maria Staicu, and B Duncan X Lascelles. The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats. *PloS one*, 12(1):e0169576, 2017.

[GF02] Charles J Golden and Shawna M Freshwater. *Stroop color and word test: A manual for clinical and experimental uses 2nd edition*. Stoelting Chicago, 2002.

[GKC+13] Patricia A Ganz, Lorna Kwan, Steven A Castellon, Amy Oppenheim, Julienne E Bower, Daniel HS Silverman, Steve W Cole, Michael R Irwin, Sonia Ancoli-Israel, and Thomas R Belin. Cognitive complaints after breast cancer treatments:

examining the relationship with neuropsychological test performance. *JNCI: Journal of the National Cancer Institute*, 105(11):791–801, 2013.

[GLJR16] Jeff Goldsmith, Xinyue Liu, Judith Jacobson, and Andrew Rundle. New insights into activity patterns in children, found using functional data analyses. *Medicine and science in sports and exercise*, 48(9):1723, 2016.

[GSE+99] Julie Akiko Gladsjo, Catherine C Schuman, Jovier D Evans, Guerry M Peavy, S Walden Miller, and Robert K Heaton. Norms for letter and category fluency: demographic corrections for age, education, and ethnicity. *Assessment*, 6(2):147–178, 1999.

[HAS+12] Elham Hedayati, Hassan Alinaghizadeh, Anna Schedin, Håkan Nyman, and Maria Albertsson. Effects of adjuvant treatment on cognitive function in women with early breast cancer. *European Journal of Oncology Nursing*, 16(3):315–322, 2012.

[HBK+13] S.D. Herrmann, T.V. Barreira, M. Kang, et al. How many hours are enough? accelerometer wear time may provide bias in daily activity estimates. *J Phys Act Health*, 10:742–9, 2013.

[HBK+14] S.D. Herrmann, T.V. Barreira, M. Kang, et al. Impact of accelerometer wear time on physical activity data: a NHANES semisimulation data approach. *Br J Sports Med*, 48:278–82, 2014.

[HKB11] J. Honaker, G. King, and M. Blackwell. Amelia ii: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[HNS+16] A. Hickey, J. Newham, M.M. Slawinska, D. Kwasnicka, S. McDonald, S. Del Din, F.F. Sniehotta, P.A. Davis, and A. Godfrey. Estimating cut points: A simple method for new wearables. *Maturitas.*, 83:78–82, 2016.

[HRP+15] Sheau-Yan Ho, Kelly J Rohan, Justin Parent, Felice A Tager, and Paula S McKinley. A longitudinal study of depression, fatigue, and sleep disturbances as a symptom cluster in women with breast cancer. *Journal of pain and symptom management*, 49(4):707–715, 2015.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Berlin Heidelberg, 2001.

[KBB+09] M. Kang, D.R. Bassett, T.V. Barreira, et al. How many days are enough? a study of 365 days of pedometer monitoring. *Res Q Exerc Sport*, 80:445–53, 2009.

[KM14] T.R. Katapally and N. Muhajarine. Towards uniform accelerometry analysis: a standardization methodology to minimize measurement bias due to systematic accelerometer wear-time variation. *J Sports Sci Med*, 13:379–86, 2014.

[KRB+09] M. Kang, D.A. Rowe, T.V. Barreira, et al. Individual information-centered approach for handling physical activity missing data. *Res Q Exerc Sport*, 80:131–7, 2009.

[KTIH00] Susan K Kongs, Laetitia L Thompson, Grant L Iverson, and Robert K Heaton. Wisconsin card sorting test-64 card version (wcst-64). *Odessa, FL: Psychological Assessment Resources*, 2000.

[KWB+13] Ling Chun Kong, Pierre-Henri Wuillemin, Jean-Philippe Bastard, Nataliya Sokolovska, Sophie Gougis, Soraya Fellahi, Froogh Darakhshan, Dominique Bonnefont-Rousselot, Randa Bittar, Joel Dore, et al. Insulin resistance and inflammation predict kinetic body weight changes in response to dietary weight loss and maintenance in overweight and obese subjects by using a bayesian network approach. *The American journal of clinical nutrition*, 98(6):1385–1394, 2013.

[LAC+14] Jennifer A Ligibel, Catherine M Alfano, Kerry S Courneya, Wendy Demark-Wahnefried, Robert A Burger, Rowan T Chlebowski, Carol J Fabian, Ayca Gucalp, Dawn L Hershman, Melissa M Hudson, et al. American society of clinical oncology position statement on obesity and cancer. *Journal of clinical oncology*, 32(31):3568–3574, 2014.

[LFN+09] Lianqi Liu, Lavinia Fiorentino, Loki Natarajan, Barbara A Parker, Paul J Mills, Georgia Robins Sadler, Joel E Dimsdale, Michelle Rissling, Feng He, and Sonia Ancoli-Israel. Pre-treatment symptom cluster in breast cancer patients is associated with worse sleep, fatigue and depression during chemotherapy. *Psycho-Oncology*, 18(2):187–194, 2009.

[LFN+16] Tran Le, Shirley W Flatt, Loki Natarajan, Bilge Pakiz, Elizabeth L Quintana, Dennis D Heath, Brinda K Rana, and Cheryl L Rock. Effects of diet composition and insulin resistance status on plasma lipid levels in a weight loss intervention in women. *Journal of the American Heart Association*, 5(1):e002771, 2016.

[LKKC16] Haocheng Li, Sarah Kozey Keadle, Victor Kipnis, and Raymond J Carroll. Longitudinal functional additive model with continuous proportional outcomes for physical activity data. *Stat*, 5(1):242–250, 2016.

[LKKS+15] Haocheng Li, Sarah Kozey Keadle, John Staudenmayer, Houssein Assaad, Jianhua Z Huang, and Raymond J Carroll. Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics*, 16(4):754–771, 2015.

[LS14] I.M. Lee and E.J. Shiroma. Using accelerometers to measure physical activity in large-scale epidemiologic studies: Issues and challenges. *Br J Sports Med.*, 48(3):197201, 2014.

[LS16] Paul D Loprinzi and Eveleen Sng. The effects of objectively measured sedentary behavior on all-cause mortality in a national sample of adults with diabetes. *Preventive medicine*, 86:55–57, 2016.

[LSC14] Haocheng Li, John Staudenmayer, and Raymond J Carroll. Hierarchical functional data with mixed continuous and binary measurements. *Biometrics*, 70(4):802–811, 2014.

[LSW+18] Haocheng Li, John Staudenmayer, Tianying Wang, Sarah Kozey Keadle, and Raymond J Carroll. Three-part joint modeling methods for complex functional data mixed with zero-and-one–inflated proportions and zero-inflated continuous outcomes with skewness. *Statistics in medicine*, 37(4):611–626, 2018.

[M+15] Gina Merchant et al. Accelerometer-measured sedentary time among hispanic adults: Results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Preventive Medicine Reports*, 2:845–853, 2015.

[MAC+06a] Jeffrey S Morris, Cassandra Arroyo, Brent A Coull, Louise M Ryan, Richard Herrick, and Steven L Gortmaker. Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association*, 101(476):1352–1364, 2006.

[MAC+06b] J.S. Morris, C. Arroyo, B.A. Coull, L.M. Ryan, R. Herrick, and S.L. Gortmaker. Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study. *J Am Stat Assoc.*, 101(476):1352–1364, 2006.

[Mas04] Mary Jane Massie. Prevalence of depression in patients with cancer. *JNCI Monographs*, 2004(32):57–71, 2004.

[MC06] Jeffrey S Morris and Raymond J Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199, 2006.

[MLF+17] Caiqin Mo, Hailong Lin, Fangmeng Fu, Lin Lin, Jie Zhang, Meng Huang, Chuan Wang, Yunjing Xue, Qing Duan, Weiwen Lin, et al. Chemotherapy-induced changes of cerebral activity in resting-state functional magnetic resonance imaging and cerebral white matter in diffusion tensor imaging. *Oncotarget*, 8(46):81273, 2017.

[N+04] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[NP15] Katharina Nimptsch and Tobias Pischon. Body fatness, related biomarkers and cancer risk: an epidemiological perspective. *Hormone molecular biology and clinical investigation*, 22(2):39–51, 2015.

[NSL13] R Nagarajan, M Scutari, and S Lèbre. Bayesian networks in r, use r, 2013.

[O'D04] J.F. O'Donnell. Insomnia in cancer patients. *Clin Cornerstone*, 6(Suppl 1D):S6–14, 2004.

[PAH+08] S.A. Prince, K.B. Adamo, M.E. Hamel, et al. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act*, 5:56, 2008.

[Par14] Timo Partonen. Obesity= physical activity+ dietary intake+ sleep stages+ light exposure. *Annals of medicine*, 46(5):245–246, 2014.

[PBD+15] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and nonlinear mixed effects models*, 2015. R package version 3.1-122.

[PCH+13] Ruth E Patterson, Graham A Colditz, Frank B Hu, Kathryn H Schmitz, Rexford S Ahima, Ross C Brownson, Kenneth R Carson, Jorge E Chavarro, Lewis A Chodosh, Sarah Gehlert, et al. The 2011–2016 transdisciplinary research on energetics and cancer (trec) initiative: rationale and design. *Cancer Causes & Control*, 24(4):695–704, 2013.

[PDRB14] Henning Palmefors, Smita DuttaRoy, Bengt Rundqvist, and Mats Borjesson. The effect of physical activity or exercise on key biomarkers in atherosclerosis–a systematic review. *Atherosclerosis*, 235(1):150–161, 2014.

[Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge university press, 2009.

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[PMN+15] R.E. Patterson, C.R. Marinac, L. Natarajan, S.J. Hartman, L. Cadmus-Bertram, S.W. Flatt, H. Li, B. Parker, J. Oratowski-Coleman, A. Villaseor, S. Godbole, and J. Kerr. Recruitment strategies, design, and participant characteristics in a trial of weight-loss and metformin in breast cancer survivors. *Contemp Clin Trials*, 2015.

[PMN+16] Ruth E Patterson, Catherine R Marinac, Loki Natarajan, Sheri J Hartman, Lisa Cadmus-Bertram, Shirley W Flatt, Hongying Li, Barbara Parker, Jesica Oratowski-Coleman, Adriana Villaseñor, et al. Recruitment strategies, design, and participant characteristics in a trial of weight-loss and metformin in breast cancer survivors. *Contemporary clinical trials*, 47:64–71, 2016.

[Rad77] Lenore Sawyer Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401, 1977.

[RS05] JO Ramsay and BW Silverman. Springer series in statistics. In *Functional data analysis*. Springer, 2005.

[RW93] RM Reitan and D Wolfson. The halstead-reitan neuropsychological test battery: Theory and clinical applications. 1993.

[SBB+14] S.D. Sanford, J.L. Beaumont, Z. Butt, J.J. Sweet, D. Cella, and L.I. Wagner. Prospective longitudinal evaluation of a symptom cluster in breast cancer. *J Pain Symptom Manage*, 47(4):721–30, 2014.

[Sch00] J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 2000.

[SFT+13] E.J. Shiroma, P.S. Freedson, S.G. Trost, et al. Patterns of accelerometer-assessed sedentary behavior in older women. *JAMA*, 310:2562–3, 2013.

[SGD+17] Francesco Sera, Lucy J Griffiths, Carol Dezateux, Marco Geraci, and Mario Cortina-Borja. Using functional data analysis to understand daily activity levels and patterns in primary school-aged children: Cross-sectional analysis of a uk-wide study. *PLoS one*, 12(11):e0187677, 2017.

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[SMG+05] P.W. Sullivan, E.H. Morrato, V. Ghushchyan, et al. Obesity, inactivity, and the prevalence of diabetes and diabetes-related cardiovascular comorbidities in the U.S., 2000-2002. *Diabetes Care*, 28:1599–603, 2005.

[SMHJ98] Kevin D Stein, Staci C Martin, Danette M Hann, and Paul B Jacobsen. A multidimensional measure of fatigue for use with cancer patients. *Cancer practice*, 6(3):143–152, 1998.

[SML+09] Winnie KW So, Gene Marsh, WM Ling, FY Leung, Joe CK Lo, Maggie Yeung, and George KH Li. The symptom cluster of fatigue, pain, anxiety, and depression and the effect on the quality of life of women receiving treatment for breast cancer: a multicenter study. In *Oncology nursing forum*, volume 36, 2009.

[SRJL+15] M1. Sani, R. Refinetti, G. Jean-Louis, S.R. Pandi-Perumal, R.A. Durazo-Arvizu, L.R. Dugas, R. Kafensztok, P. Bovet, T.E. Forrester, E.V. Lambert, J. Plange-Rhule, and A. Luke. Daily activity patterns of 2316 men and women from five countries differing in socioeconomic development. *Chronobiol Int.*, 32(5):650–6, 2015.

[SS00] J.F. Sallis and B.E. Saelens. Assessment of physical activity by self-report: status, limitations, and future directions. *Res Q Exerc Sport*, 71:S1–14, 2000.

[SSB+00] AM1. Swartz, S.J. Strath, D.R. Jr. Bassett, W.L. O'Brien, G.A. King, and B.E. Ainsworth. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc.*, 32:S450–6, 2000.

[SZC12] J. Staudenmayer, W. Zhu, and D.J. Catellier. Statistical considerations in the analysis of accelerometry-based activity monitor data. *Med Sci Sports Exerc*, 44:S61–7, 2012.

[SZCG15] Haochang Shou, Vadim Zipunnikov, Ciprian M Crainiceanu, and Sonja Greven. Structured functional principal component analysis. *Biometrics*, 71(1):247–257, 2015.

[TB07] Jesper Tegnér and Johan Björkegren. Perturbations to uncover gene networks. *TRENDS in Genetics*, 23(1):34–41, 2007.

[TBD+08] R.P. Troiano, D. Berrigan, K.W. Dodd, et al. Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc*, 40:181–8, 2008.

[TMP05] S.G. Trost, K.L. McIver, and R.R. Pate. Conducting accelerometer-based activity assessments in field-based research. *Med Sci Sports Exerc*, 37:S531–43, 2005.

[VB02] Harri Vainio and Franca Bianchini. *Weight control and physical activity*. Number 6. Iarc, 2002.

[vLWV+12] S.W. van Landingham, J.R. Willis, S. Vitale, et al. Visual field loss and accelerometer-measured physical activity in the United States. *Ophthalmology*, 119:2486–92, 2012.

[VPR+07] Marion Verduijn, Niels Peek, Peter MJ Rosseel, Evert de Jonge, and Bas AJM de Mol. Prognostic bayesian networks: I: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, 40(6):609–618, 2007.

[VRW+09] Katharine J Vearncombe, Margaret Rolfe, Margaret Wright, Nancy A Pachana, Brooke Andrew, and Geoffrey Beadle. Predictors of cognitive decline after chemotherapy in breast cancer patients. *Journal of the International Neuropsychological Society*, 15(6):951–962, 2009.

[Was00] Larry Wasserman. Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000.

[WBH09] Kimberly K Wielgus, Ann M Berger, and Melody Hertzog. Predictors of fatigue 30 days after completing anthracycline plus taxane adjuvant chemotherapy for breast cancer. In *Oncology nursing forum*, volume 36, 2009.

[WEA+12] E.G. Wilmot, C.L. Edwardson, F.A. Achana, M.J. Davies, T. Gorely, L.J. Gray, K. Khunti, T. Yates, and S.J. Biddle. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. *Diabetologia*, 55(11):2895–905, 2012.

[Wec14] David Wechsler. Wechsler adult intelligence scale–fourth edition (wais–iv). *San Antonio, Texas: Psychological Corporation*, 2014.

[WEV+05] D.S. Ward, K.R. Evenson, A. Vaughn, et al. Accelerometer use in physical activity: best practices and research recommendations. *Med Sci Sports Exerc*, 37:S582–8, 2005.

[WH15] Horng-Shiuann Wu and Janet K Harden. Symptom burden and quality of life in survivorship: a review of the literature. *Cancer nursing*, 38(1):E29–E54, 2015.

[WLE+97] Terri E Weaver, Andréa M Laizner, Lois K Evans, Greg Maislin, Deepak K Chugh, Kerry Lyon, Philip L Smith, Alan R Schwartz, Susan Redline, Allan I Pack, et al. An instrument to measure functional status outcomes for disorders of excessive sleepiness. *Sleep*, 20(10):835–843, 1997.

[WVHS+15] D.K. Wilson, M.L. Van Horn, E.R. Siceloff, et al. The results of the Positive Action for Today's Health" (PATH) trial for increasing walking and physical activity in underserved African-American communities. *Ann Behav Med*, 49:398–410, 2015.

[XHS+14] Luo Xiao, Lei Huang, Jennifer A Schrack, Luigi Ferrucci, Vadim Zipunnikov, and Ciprian M Crainiceanu. Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics*, 16(2):352–367, 2014.

[XHS+15] L. Xiao, L. Huang, J.A. Schrack, L. Ferrucci, V. Zipunnikov, and C.M. Crainiceanu. Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics.*, 16(2):352–67, 2015.

[XKHM14] Qian Xiao, Sarah K Keadle, Albert R Hollenbeck, and Charles E Matthews. Sleep duration and total and cause-specific mortality in a large us cohort: interrelationships with physical activity, sedentary behavior, and body mass index. *American journal of epidemiology*, 180(10):997–1006, 2014.

[YJC+09] D.R. Young, G.J. Jerome, C. Chen, et al. Patterns of physical activity among overweight and obese adults. *Prev Chronic Dis*, 6:A90, 2009.

[YRL14] Changwon Yoo, Luis Ramirez, and Juan Liuzzi. Big data analysis using modern statistical and machine learning methods in medicine. *International neurourology journal*, 18(2):50, 2014.

[ZBH+17] Jamie M Zeitzer, Terri Blackwell, Andrew R Hoffman, Steve Cummings, Sonia Ancoli-Israel, Katie Stone, and Osteoporotic Fractures in Men (MrOS) Study Research Group. Daily patterns of accelerometer activity predict changes in sleep, cognition, and mortality in older men. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, page glw250, 2017.

[ZSAK+06] Lei Zhang, Dimitris Samaras, Nelly Alia-Klein, Nora Volkow, and Rita Goldstein. Modeling neuronal interactivity using dynamic bayesian networks. In *Advances in neural information processing systems*, pages 1593–1600, 2006.